1. Syllabus
    1. Letter to Student
    2. Contact Information
    3. Resources
    4. Purpose of the Course
    5. Course Description
    6. Calendar
    7. Grading Procedures
2. Chapter 1: Signals and Systems
    1. Signal Classifications and Properties
    2. System Classifications and Properties
    3. m04 - Theorems on the Fourier Series
    4. m05 - The Fourier Transform
    5. Review of Probability Theory
    6. Introduction to Stochastic Processes
    7. Second-order Description
    8. Gaussian Processes
    9. White and Coloured Processes
    10. Linear Filtering
3. Chapter 2: Source Coding
    1. Information Theory and Coding
    2. Entropy
    3. Source Coding
    4. Huffman Coding
4. Chapter 3: Communication over AWGN Channels
    1. Data Transmission and Reception
    2. Signalling
    3. Geometric Representation of Modulation Signals
    4. Demodulation and Detection
    5. Demodulation
    6. Detection by Correlation

    7. Examples of Correlation Detection
    8. Matched Filters
    9. Examples with Matched Filters
    10. Performance Analysis of Binary Orthogonal Signals with Correlation
    11. Performance Analysis of Orthogonal Binary Signals with Matched Filters
    12. Carrier Phase Modulation
    13. Carrier Frequency Modulation
    14. Differential Phase Shift Keying
5. Chapter 4: Communication over Band-limitted AWGN Channel
    1. Digital Transmission over Baseband Channels
    2. Introduction to ISI
    3. Pulse Amplitude Modulation Through Bandlimited Channel
    4. Precoding and Bandlimited Signals
    5. Pulse Shaping to Reduce ISI
    6. Two Types of Error-Performance Degradation
    7. Eye Pattern
    8. Transversal Equalizer
    9. Decision Feedback Equalizer
    10. Adaptive Equalization
6. Chapter 5: Channel Coding
    1. Channel Capacity
    2. Mutual Information
    3. Typical Sequences
    4. Shannon's Noisy Channel Coding Theorem
    5. Channel Coding
    6. Convolutional Codes
7. Chapter 6: Communication over Fading Channels
    1. Fading Channel

2. Characterizing Mobile-Radio Propagation
3. Large-Scale Fading
4. Small-Scale Fading
5. Signal Time-Spreading
6. Mitigating the Degradation Effects of Fading
7. Mitigation to Combat Frequency-Selective Distortion
8. Mitigation to Combat Fast-Fading Distortion
9. Mitigation to Combat Loss in SNR
10. Diversity Techniques
11. Diversity-Combining Techniques
12. Modulation Types for Fading Channels
13. The Role of an Interleaver
14. The Viterbi Equalizer as Applied to GSM
15. The Rake Receiver Applied to Direct-Sequence Spread-Spectrum (DS/SS) Systems

Letter to Student

To the Student:

This course and this Student Manual reflect a collective effort by your instructor, the Vietnam Education Foundation, the Vietnam Open Courseware (VOCW) Project and faculty colleagues within Vietnam and the United States who served as reviewers of drafts of this Student Manual. This course is an important component of our academic program. Although it has been offered for many years, this latest version represents an attempt to expand the range of sources of information and instruction so that the course continues to be up-to-date and the methods well suited to what is to be learned.

This Student Manual is designed to assist you through the course by providing specific information about student responsibilities including requirements, timelines and evaluations.

You will be asked from time-to-time to offer feedback on how the Student Manual is working and how the course is progressing. Your comments will inform the development team about what is working and what requires attention. Our goal is to help you learn what is important about this particular field and to eventually succeed as a professional applying what you learn in this course.

Thank you for your cooperation.

Tuan Do-Hong.

Contact Information

Faculty Information: Department of Telecommunications Engineering, Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology

Instructor: Dr.-Ing. Tuan Do-Hong

Office Location: Ground floor, B3 Building

Phone: +84 (0) 8 8654184

Email: do-hong@hcmut.edu.vn

Office Hours: 9:00 am – 5:00 pm

Assistants:

Office Location: Ground floor, B3 Building

Phone: +84 (0) 8 8654184

Email:

Office Hours: 9:00 am – 5:00 pm

Lab sections/support:

Resources

Connexions: http://cnx.org/

MIT's OpenCourseWare: http://ocw.mit.edu/index.html

Computer resource: Matlab and Simulink

Textbook(s):

Required:

[1] Bernard Sklar, **Digital Communications: Fundamentals and Applications**, 2nd edition, 2001, Prentice Hall.

Recommended:

[2] John Proakis, **Digital Communications**, 4th edition, 2001, McGraw-Hill.

[3] Bruce Carlson et al., **Communication Systems: An Introduction to Signals and Noise in Electrical Communication**, 4th edition, 2001, McGraw-Hill.

[4] Rogger E. Ziemer, Roger W. Peterson, **Introduction to Digital Communication**, 2nd edition, 2000, Prenctice Hall.

Purpose of the Course

Title: Principles of Digital Communications

Credits: 3 (4 hours/week, 15 weeks/semester)

Course Rationale:

Wireless communication is fundamentally the art of communicating information without wires. In principle, wireless communication encompasses any number of techniques including underwater acoustic communication, radio communication, and satellite communication, among others. The term was coined in the early days of radio, fell out of fashion for about fifty years, and was rediscovered during the cellular telephony revolution. Wireless now implies communication using electromagnetic waves - placing it within the domain of electrical engineering. Wireless communication techniques can be classified as either analog or digital. The first commercial systems were analog including AM radio, FM radio, television, and first generation cellular systems. Analog communication is gradually being replaced with digital communication. The fundamental difference between the two is that in digital communication, the source is assumed to be digital. Every major wireless system being developed and deployed is built around digital communication including cellular communication, wireless local area networking, personal area networking, and high-definition television. Thus this course will focus on digital wireless communication.

This course is a required core course in communications engineering which introduces principles of digital communications while reinforcing concepts learned in analog communications systems. It is intended to provide a comprehensive coverage of digital communication systems for last year undergraduate students, first year graduate students and practicing engineers.

Pre-requisites: **Communication Systems**. Thorough knowledge of **Signals and Systems**, **Linear Algebra**, **Digital Signal Processing**, and **Probability Theory and Stochastic Processes** is essential.

Course Description

This course explores elements of the theory and practice of digital communications. The course will 1) model and study the effects of channel impairments such as distortion, noise, interference, and fading, on the performance of communication systems; 2) introduce signal processing, modulation, and coding techniques that are used in digital communication systems. The concepts/ tools are acquired in this course:

Signals and Systems

Classification of signals and systems

Orthogonal functions, Fourier series, Fourier transform

Spectra and filtering

Sampling theory, Nyquist theorem

Random processes, autocorrelation, power spectrum

Systems with random input/output

Source Coding

Elements of compression, Huffman coding

Elements of quantization theory

Pulse code Modulation (PCM) and variations

Rate/bandwidth calculations in communication systems

Communication over AWGN Channels

Signals and noise, Eb/N0

Receiver structure, demodulation and detection

Correlation receiver and matched filter

Detection of binary signals in AWGN

Optimal detection for general modulation

Coherent and non-coherent detection

Communication over Band-limited AWGN Channel

ISI in band-limited channels

Zero-ISI condition: the Nyquist criterion

Raised cosine filters

Partial response signals

Equalization using zero-forcing criterion

Channel Coding

Types of error control

Block codes

Error detection and correction

Convolutional codes and the Viterbi algorithm

Communication over Fading Channel

Fading channels

Characterizing mobile-radio propagation

Signal Time-Spreading

Mitigating the effects of fading

Application of Viterbi equalizer in GSM system

Application of Rake receiver in CDMA system

Calendar

Week 1: Overview of signals and spectra

Week 2: Source coding

Week 3: Receiver structure, demodulation and detection

Week 4: Correlation receiver and matched filter. Detection of binary signals in AWGN

Week 5: Optimal detection for general modulation. Coherent and non-coherent detection (I)

Week 6: Coherent and non-coherent detection (II)

Week 7: ISI in band-limited channels. Zero-ISI condition: the Nyquist criterion

Week 8: Mid-term exam

Week 9: Raised cosine filters. Partial response signals

Week 10: Channel equalization

Week 11: Channel coding. Block codes

Week 12: Convolutional codes

Week 13: Viterbi algorithm

Week 14: Fading channel. Characterizing mobile-radio propagation

Week 15: Mitigating the effects of fading

Week 16: Applications of Viterbi equalizer and Rake receiver in GSM and CDMA systems

Week 17: Final exam

Grading Procedures

Homework/Participation/Exams:

- Homework and Programming Assignments
- Midterm Exam
- Final Exam

Homework and programming assignments will be given to test student's knowledge and understanding of the covered topics. Homework and programming assignments will be assigned frequently throughout the course and will be due in the time and place indicated on the assignment. Homework and programming assignments must be individually done by each student without collaboration with others. No late homework will be allowed.

There will be in-class mid-term and final exams. The mid-term exam and the final exam will be time-limited to 60 minutes and 120 minutes, respectively. They will be closed book and closed notes. It is recommend that the students practice working problems from the book, example problems, and homework problems.

Participation: Question and discussion in class are encouraged. Participation will be noted.

Grades for this course will be based on the following weighting:

- Homework and In-class Participation: 20%
- Programming Assignments: 20%
- Mid-term Exam: 20%
- Final Exam: 40%

Signal Classifications and Properties
Describes various classifications of signals.

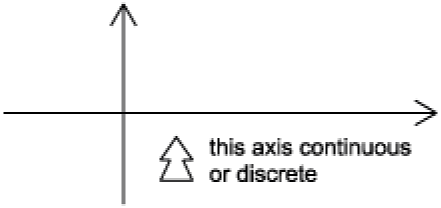
Introduction

This module will begin our study of signals and systems by laying out some of the fundamentals of signal classification. It is essentially an introduction to the important definitions and properties that are fundamental to the discussion of signals and systems, with a brief discussion of each.

Classifications of Signals

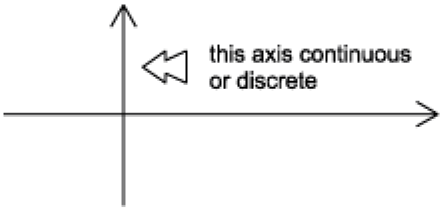
Continuous-Time vs. Discrete-Time

As the names suggest, this classification is determined by whether or not the time axis is **discrete** (countable) or **continuous** ([link]). A continuous-time signal will contain a value for all real numbers along the time axis. In contrast to this, a discrete-time signal, often created by sampling a continuous signal, will only have values at equally spaced intervals along the time axis.



Analog vs. Digital

The difference between **analog** and **digital** is similar to the difference between continuous-time and discrete-time. However, in this case the difference involves the values of the function. Analog corresponds to a continuous set of possible function values, while digital corresponds to a discrete set of possible function values. An common example of a digital signal is a binary sequence, where the values of the function can only be one or zero.



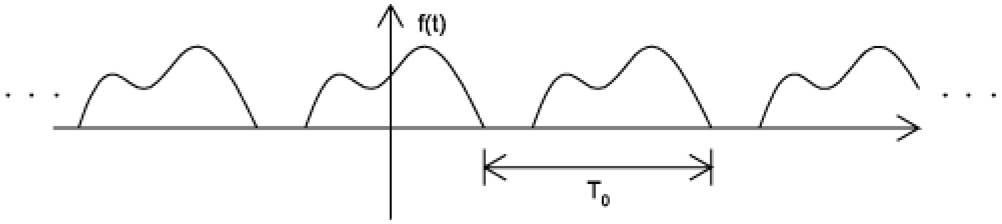
Periodic vs. Aperiodic

Periodic signals repeat with some **period** T, while aperiodic, or nonperiodic, signals do not ([link]). We can define a periodic function through the following mathematical expression, where t can be any number and T is a positive constant:

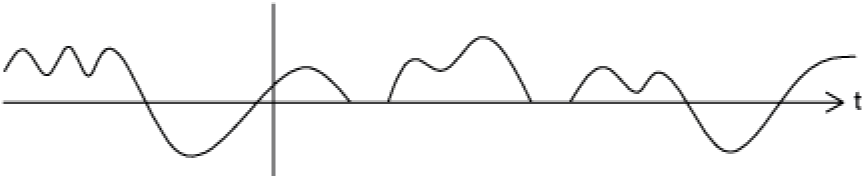
Equation:

$$f(t)=f(t+T)$$

fundamental period of our function, $f(t)$, is the smallest value of T that the still allows [link] to be true.



A periodic signal with period T_0



An aperiodic signal

Finite vs. Infinite Length

Another way of classifying a signal is in terms of its length along its time axis. Is the signal defined for all possible values of time, or for only certain values of time? Mathematically speaking, $f(t)$ is a **finite-length signal** if it is **defined** only over a finite interval

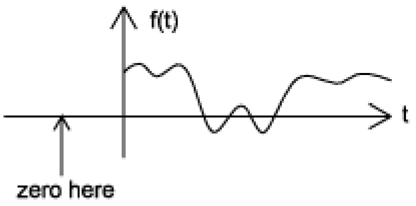
$$t_1 < t < t_2$$

where $t_1 < t_2$. Similarly, an **infinite-length signal**, $f(t)$, is defined for all values:

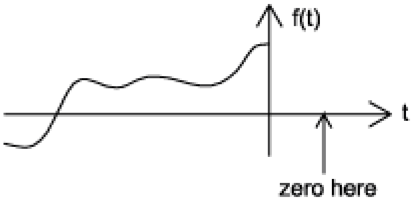
$$-\infty < t < \infty$$

Causal vs. Anticausal vs. Noncausal

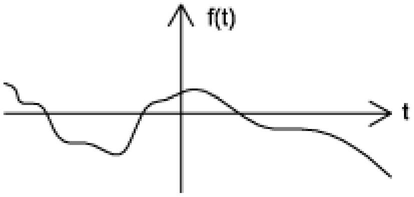
Causal signals are signals that are zero for all negative time, while **anticausal** are signals that are zero for all positive time. **Noncausal** signals are signals that have nonzero values in both positive and negative time ([link]).



A causal signal



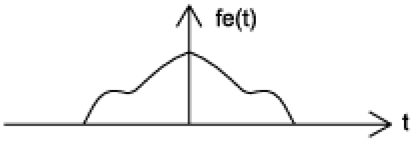
An anticausal signal



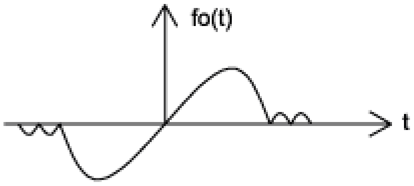
A noncausal signal

Even vs. Odd

An **even signal** is any signal f such that $f(t) = f(-t)$. Even signals can be easily spotted as they are **symmetric** around the vertical axis. An **odd signal**, on the other hand, is a signal f such that $f(t) = -f(-t)$ ([link]).



An even signal



An odd signal

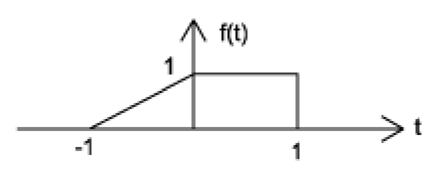
Using the definitions of even and odd signals, we can show that any signal can be written as a combination of an even and odd signal. That is, every signal has an odd-even decomposition. To demonstrate this, we have to look no further than a single equation.

Equation:

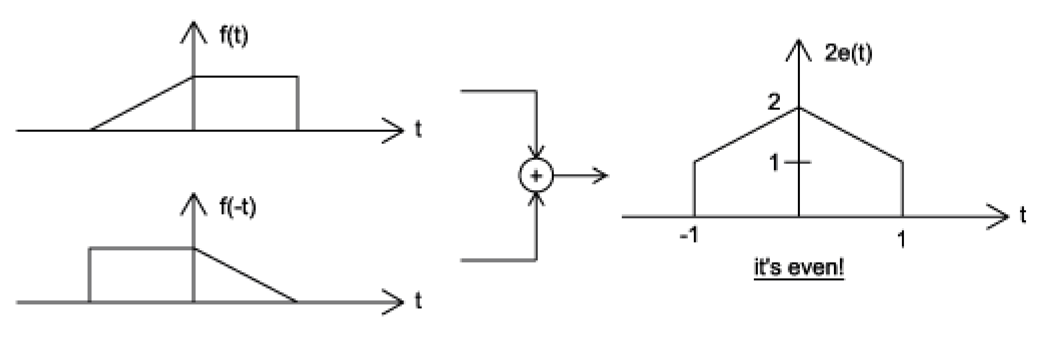
$$f(t)=\frac{1}{2}\left(f(t)+f(-t)\right)+\frac{1}{2}\left(f(t)-f(-t)\right)$$

By multiplying and adding this expression out, it can be shown to be true. Also, it can be shown that $f(t)+f(-t)$ fulfills the requirement of an even function, while $f(t)-f(-t)$ fulfills the requirement of an odd function ([link]).

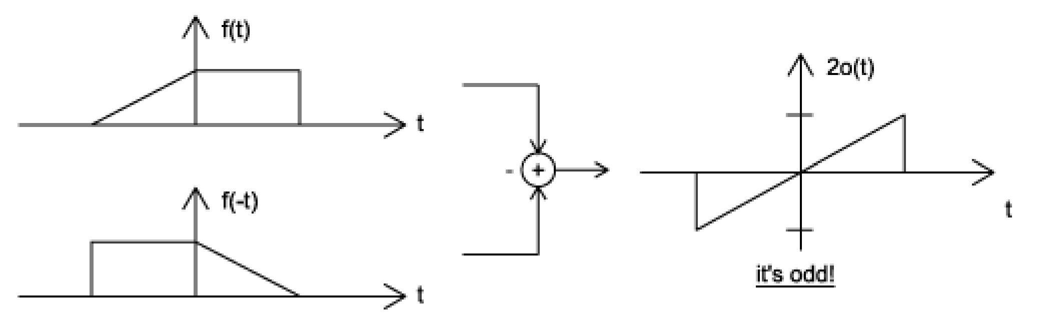
Example:



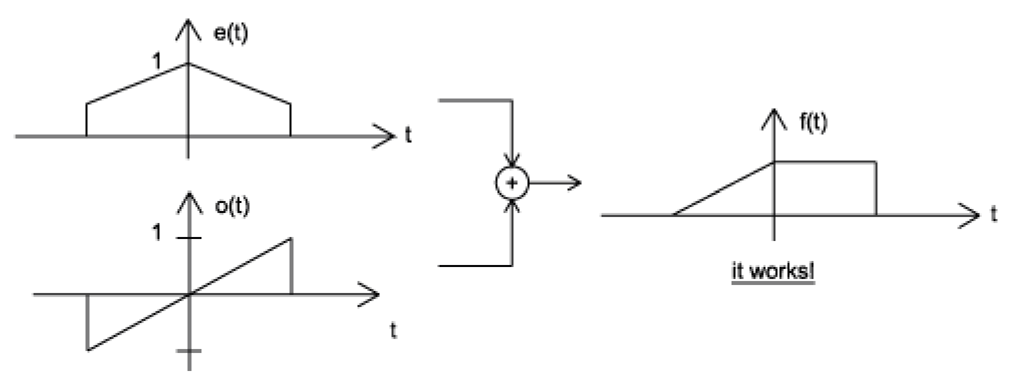
The signal we will decompose using odd-even decomposition



Even part: $e(t) = \frac{1}{2}\left(f(t) + f(-t)\right)$



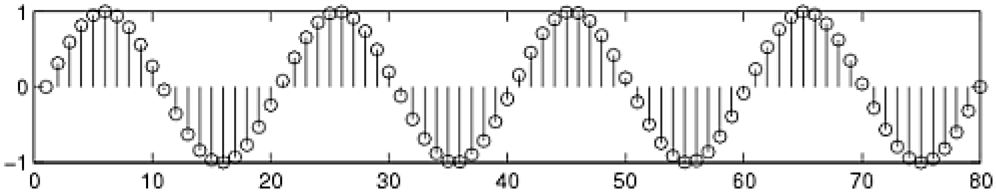
Odd part: $o(t) = \frac{1}{2}\left(f(t) - f(-t)\right)$



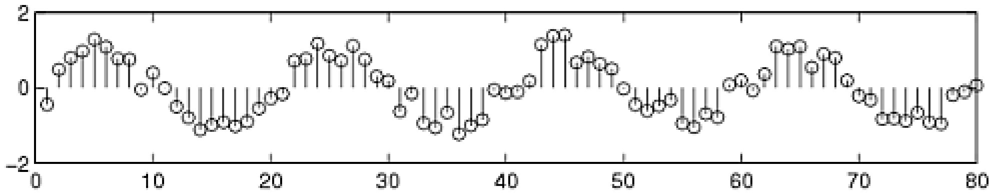
Check: $e(t) + o(t) = f(t)$

Deterministic vs. Random

A **deterministic signal** is a signal in which each value of the signal is fixed, being determined by a mathematical expression, rule, or table. On the other hand, the values of a **random signal** are not strictly defined, but are subject to some amount of variability.



Deterministic Signal



Random Signal

Example:
Consider the signal defined for all real t described by
Equation:

$$f(t) = \begin{cases} \sin(2\pi t)/t & t \geq 1 \\ 0 & t < 1 \end{cases}$$

This signal is continuous time, analog, aperiodic, infinite length, causal, neither even nor odd, and, by definition, deterministic.

Signal Classifications Summary

This module describes just some of the many ways in which signals can be classified. They can be continuous time or discrete time, analog or digital, periodic or aperiodic, finite or infinite, and deterministic or random. We can also divide them based on their causality and symmetry properties.

System Classifications and Properties
Describes various classifications of systems.

Introduction

In this module some of the basic classifications of systems will be briefly introduced and the most important properties of these systems are explained. As can be seen, the properties of a system provide an easy way to distinguish one system from another. Understanding these basic differences between systems, and their properties, will be a fundamental concept used in all signal and system courses. Once a set of systems can be identified as sharing particular properties, one no longer has to reprove a certain characteristic of a system each time, but it can simply be known due to the the system classification.

Classification of Systems

Continuous vs. Discrete

One of the most important distinctions to understand is the difference between discrete time and continuous time systems. A system in which the input signal and output signal both have continuous domains is said to be a continuous system. One in which the input signal and output signal both have discrete domains is said to be a discrete system. Of course, it is possible to conceive of signals that belong to neither category, such as systems in which sampling of a continuous time signal or reconstruction from a discrete time signal take place.

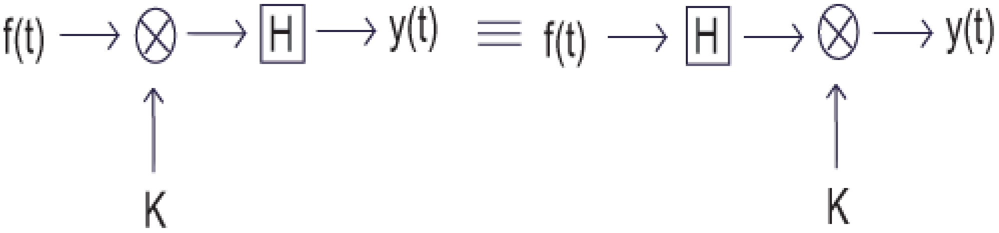
Linear vs. Nonlinear

A linear system is any system that obeys the properties of scaling (first order homogeneity) and superposition (additivity) further described below. A nonlinear system is any system that does not have at least one of these properties.

To show that a system H obeys the scaling property is to show that

Equation:

$$H(kf(t)) = kH(f(t))$$

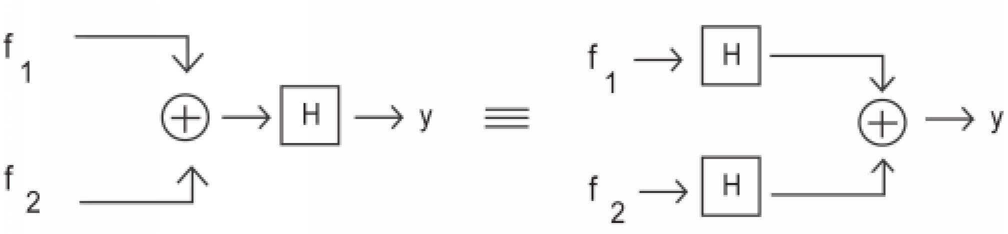


A block diagram demonstrating the scaling property of linearity

To demonstrate that a system H obeys the superposition property of linearity is to show that

Equation:

$$H(f_1(t) + f_2(t)) = H(f_1(t)) + H(f_2(t))$$



A block diagram demonstrating the superposition property of linearity

It is possible to check a system for linearity in a single (though larger) step. To do this, simply combine the first two steps to get

Equation:

$$H(k_1 f_1(t) + k_2 f_2(t)) = k_1 H(f_1(t)) + k_2 H(f_2(t))$$

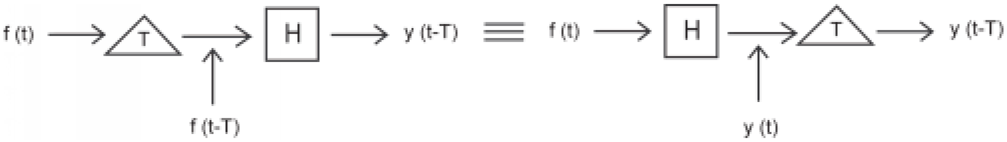
Time Invariant vs. Time Varying

A system is said to be time invariant if it commutes with the parameter shift operator defined by $S_T(f(t)) = f(t - T)$ for all T, which is to say

Equation:

$$HS_T = S_T H$$

for all real T. Intuitively, that means that for any input function that produces some output function, any time shift of that input function will produce an output function identical in every way except that it is shifted by the same amount. Any system that does not have this property is said to be time varying.

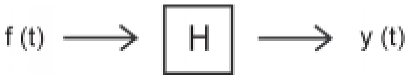


This block diagram shows what the condition for time invariance. The output is the same whether the delay is put on the input or the output.

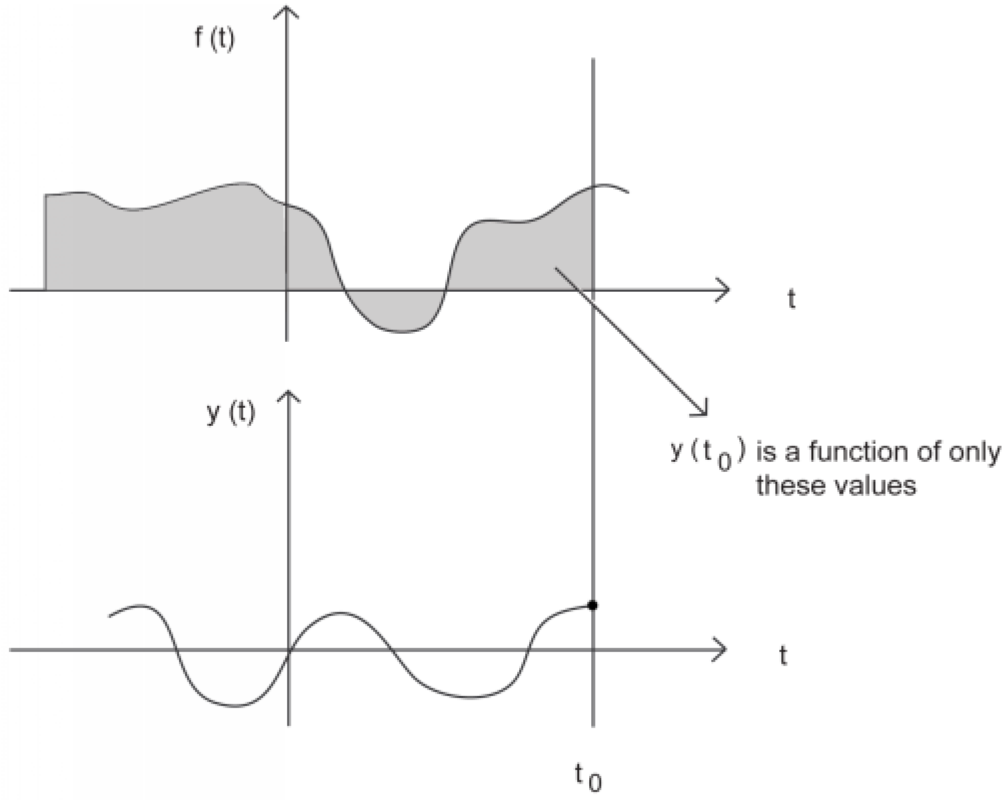
Causal vs. Noncausal

A causal system is one in which the output depends only on current or past inputs, but not future inputs. Similarly, an anticausal system is one in which the output depends only on current or future inputs, but not past inputs. Finally, a noncausal system is one in which the output depends on both past and future inputs. All "realtime" systems must be causal, since they can not have future inputs available to them.

One may think the idea of future inputs does not seem to make much physical sense; however, we have only been dealing with time as our dependent variable so far, which is not always the case. Imagine rather that we wanted to do image processing. Then the dependent variable might represent pixel positions to the left and right (the "future") of the current position on the image, and we would not necessarily have a causal system.



For a typical system to be causal...



...the output at time t_0, $y(t_0)$, can only depend on the portion of the input signal before t_0.

Stable vs. Unstable

There are several definitions of stability, but the one that will be used most frequently in this course will be bounded input, bounded output (BIBO) stability. In this context, a stable system is one in which the output is bounded if the input is also bounded. Similarly, an unstable system is one in which at least one bounded input produces an unbounded output.

Representing this mathematically, a stable system must have the following property, where $x(t)$ is the input and $y(t)$ is the output. The output must satisfy the condition

Equation:

$$|y(t)| \leq M_y < \infty$$

whenever we have an input to the system that satisfies
Equation:

$$|x(t)| \leq M_x < \infty$$

M_x and M_y both represent a set of finite positive numbers and these relationships hold for all of t. Otherwise, the system is unstable.

System Classifications Summary

This module describes just some of the many ways in which systems can be classified. Systems can be continuous time, discrete time, or neither. They can be linear or nonlinear, time invariant or time varying, and stable or unstable. We can also divide them based on their causality properties. There are other ways to classify systems, such as use of memory, that are not discussed here but will be described in subsequent modules.

Theorems on the Fourier Series

Four of the most important theorems in the theory of Fourier analysis are the inversion theorem, the convolution theorem, the differentiation theorem, and Parseval's theorem [link]. All of these are based on the orthogonality of the basis function of the Fourier series and integral and all require knowledge of the convergence of the sums and integrals. The practical and theoretical use of Fourier analysis is greatly expanded if use is made of distributions or generalized functions [link][link]. Because energy is an important measure of a function in signal processing applications, the Hilbert space of L^2 functions is a proper setting for the basic theory and a geometric view can be especially useful [link][link].

The following theorems and results concern the existence and convergence of the Fourier series and the discrete-time Fourier transform [link]. Details, discussions and proofs can be found in the cited references.

- If $f(x)$ has bounded variation in the interval $(-\pi, \pi)$, the Fourier series corresponding to $f(x)$ converges to the value $f(x)$ at any point within the interval, at which the function is continuous; it converges to the value $\frac{1}{2}[f(x+0)+f(x-0)]$ at any such point at which the function is discontinuous. At the points $\pi, -\pi$ it converges to the value $\frac{1}{2}[f(-\pi+0)+f(\pi-0)]$. [link]
- If $f(x)$ is of bounded variation in $(-\pi, \pi)$, the Fourier series converges to $f(x)$, uniformly in any interval (a, b) in which $f(x)$ is continuous, the continuity at a and b being on both sides. [link]
- If $f(x)$ is of bounded variation in $(-\pi, \pi)$, the Fourier series converges to $\frac{1}{2}[f(x+0)+f(x-0)]$, bounded throughout the interval $(-\pi, \pi)$. [link]
- If $f(x)$ is bounded and if it is continuous in its domain at every point, with the exception of a finite number of points at which it may have ordinary discontinuities, and if the domain may be divided into a finite number of parts, such that in any one of them the function is monotone; or, in other words, the function has only a finite number of

maxima and minima in its domain, the Fourier series of $f(x)$ converges to $f(x)$ at points of continuity and to $\frac{1}{2}[f(x+0)+f(x-0)]$ at points of discontinuity. [link][link]

- If $f(x)$ is such that, when the arbitrarily small neighborhoods of a finite number of points in whose neighborhood $|f(x)|$ has no upper bound have been excluded, $f(x)$ becomes a function with bounded variation, then the Fourier series converges to the value $\frac{1}{2}[f(x+0)+f(x-0)]$, at every point in $(-\pi,\pi)$, except the points of infinite discontinuity of the function, provided the improper integral $\int_{-\pi}^{\pi} f(x)dx$ exist, and is absolutely convergent. [link]
- If f is of bounded variation, the Fourier series of f converges at every point x to the value $[f(x+0)+f(x-0)]/2$. If f is, in addition, continuous at every point of an interval $I=(a,b)$, its Fourier series is uniformly convergent in I. [link]
- If $a(k)$ and $b(k)$ are absolutely summable, the Fourier series converges uniformly to $f(x)$ which is continuous. [link]
- If $a(k)$ and $b(k)$ are square summable, the Fourier series converges to $f(x)$ where it is continuous, but not necessarily uniformly. [link]
- Suppose that $f(x)$ is periodic, of period X, is defined and bounded on $[0,X]$ and that at least one of the following four conditions is satisfied: (i) f is piecewise monotonic on $[0,X]$, (ii) f has a finite number of maxima and minima on $[0,X]$ and a finite number of discontinuities on $[0,X]$, (iii) f is of bounded variation on $[0,X]$, (iv) f is piecewise smooth on $[0,X]$: then it will follow that the Fourier series coefficients may be defined through the defining integral, using proper Riemann integrals, and that the Fourier series converges to $f(x)$ at a.a.x, to $f(x)$ at each point of continuity of f, and to the value $\frac{1}{2}[f(x^-)+f(x^+)]$ at all x. [link]
- For any $1 \leq p < \infty$ and any $f \in C^p(S^1)$, the partial sums
  Equation:

$$S_n = S_n(f) = \underset{|k|\leq n}{\quad} \widehat{f}(k)e_k$$

converge to f, uniformly as $n \to \infty$; in fact, $||S_n - f||_\infty$ is bounded by a constant multiple of $n^{-p+1/2}$. [link]

The Fourier series expansion results in transforming a periodic, continuous time function, $\widetilde{x}(t)$, to two discrete indexed frequency functions, $a(k)$ and $b(k)$ that are not periodic.

m05 - The Fourier Transform

The Fourier Transform

Many practical problems in signal analysis involve either infinitely long or very long signals where the Fourier series is not appropriate. For these cases, the Fourier transform (FT) and its inverse (IFT) have been developed. This transform has been used with great success in virtually all quantitative areas of science and technology where the concept of frequency is important. While the Fourier series was used before Fourier worked on it, the Fourier transform seems to be his original idea. It can be derived as an extension of the Fourier series by letting the length increase to infinity or the Fourier transform can be independently defined and then the Fourier series shown to be a special case of it. The latter approach is the more general of the two, but the former is more intuitive [link][link].

Definition of the Fourier Transform

The Fourier transform (FT) of a real-valued (or complex) function of the real-variable t is defined by

Equation:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} \mathbb{d}t$$

giving a complex valued function of the real variable ω representing frequency. The inverse Fourier transform (IFT) is given by

Equation:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} \mathbb{d}\omega.$$

Because of the infinite limits on both integrals, the question of convergence is important. There are useful practical signals that do not have Fourier transforms if only classical functions are allowed because of problems with convergence. The use of delta functions (distributions) in both the time and frequency domains allows a much larger class of signals to be represented [link].

Examples of the Fourier Transform

Deriving a few basic transforms and using the properties allows a large class of signals to be easily studied. Examples of modulation, sampling, and others will be given.

- If $x(t)=\delta(t)$ then $X(\omega)=1$
- If $x(t)=1$ then $X(\omega)=2\pi\delta(\omega)$
- If $x(t)$ is an infinite sequence of delta functions spaced T apart, $x(t)=\sum_{n=-\infty}^{\infty}\delta(t-nT)$, its transform is also an infinite sequence of delta functions of weight $2\pi/T$ spaced $2\pi/T$ apart, $X(\omega)=2\pi\sum_{k=-\infty}^{\infty}\delta(\omega-2\pi k/T)$.
- Other interesting and illustrative examples can be found in [link][link].

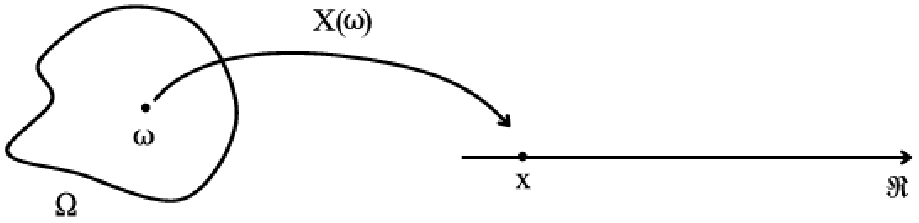
Note the Fourier transform takes a function of continuous time into a function of continuous frequency, neither function being periodic. If distribution" or delta functions" are allowed, the Fourier transform of a periodic function will be a infinitely long string of delta functions with weights that are the Fourier series coefficients.

Review of Probability Theory

The focus of this course is on digital communication, which involves transmission of information, in its most general sense, from source to destination using digital technology. Engineering such a system requires modeling both the information and the transmission media. Interestingly, modeling both digital or analog information and many physical media requires a probabilistic setting. In this chapter and in the next one we will review the theory of probability, model random signals, and characterize their behavior as they traverse through deterministic systems disturbed by noise and interference. In order to develop practical models for random phenomena we start with carrying out a random experiment. We then introduce definitions, rules, and axioms for modeling within the context of the experiment. The outcome of a random experiment is denoted by ω. The sample space Ω is the set of all possible outcomes of a random experiment. Such outcomes could be an abstract description in words. A scientific experiment should indeed be repeatable where each outcome could naturally have an associated probability of occurrence. This is defined formally as the ratio of the number of times the outcome occurs to the total number of times the experiment is repeated.

Random Variables

A random variable is the assignment of a real number to each outcome of a random experiment.



Example:
Roll a dice. Outcomes $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
ω_i = i dots on the face of the dice.
$X(\omega_i) = i$

Distributions

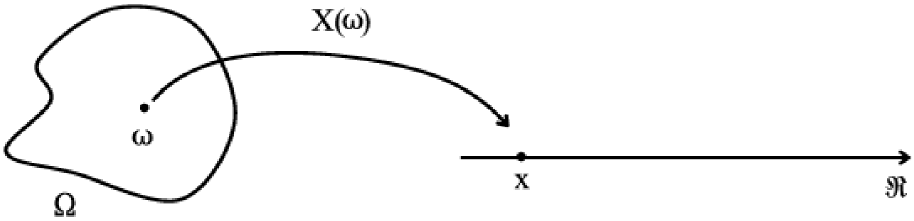
Probability assignments on intervals $a < X \leq b$

Cumulative distribution

The cumulative distribution function of a random variable X is a function $\mathrm{F}_X \; (\mathbb{R} \mapsto \mathbb{R})$ such that

Equation:

$$\begin{aligned} \mathrm{F}_X \; (b) &= \Pr[X \leq b] \\ &= \Pr[\{\omega \in \Omega | \; X(\omega) \leq b\}] \end{aligned}$$



Continuous Random Variable

A random variable X is continuous if the cumulative distribution function can be written in an integral form, or

Equation:

$$\mathrm{F}_X \; (b) = \int_{-\infty}^{b} \mathrm{f}_X \; (x) \; \mathrm{d} \; x$$

and $\mathrm{f}_X \; (x)$ is the probability density function (pdf) (e.g., $\mathrm{F}_X \; (x)$ is differentiable and $\mathrm{f}_X \; (x) = \frac{\mathrm{d}}{\mathrm{d}x}(\mathrm{F}_X \; (x)))$

Discrete Random Variable

A random variable X is discrete if it only takes at most countably many points (i.e., $\mathrm{F}_X \; (\cdot)$ is piecewise constant). The probability mass function (pmf) is defined as

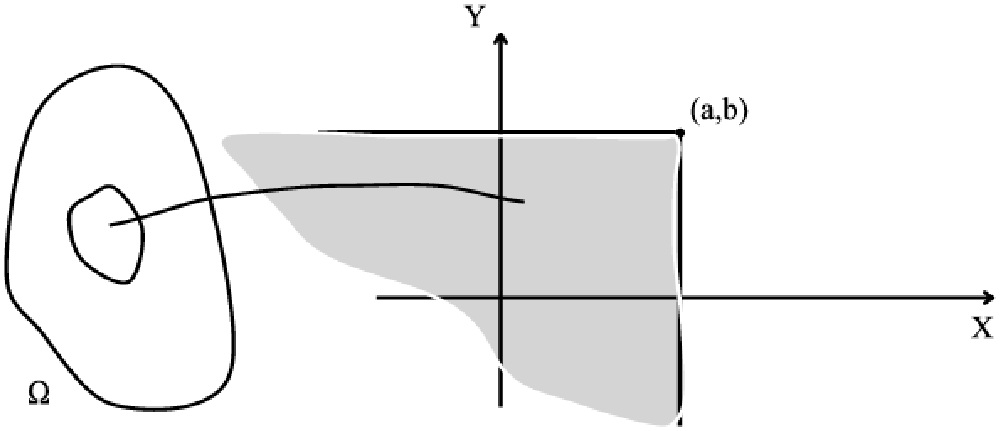
Equation:

$$\begin{aligned} \mathrm{p}_X \; (x_k) &= \Pr[X = x_k] \\ &= \mathrm{F}_X \; (x_k) - \underset{x(x \rightarrow x_k) \; \wedge \; (x < x_k)}{\text{limit}} \mathrm{F}_X \; (x) \end{aligned}$$

Two random variables defined on an experiment have joint distribution

Equation:

$$\begin{aligned}\mathrm{F}_{X,,,Y}\ (a,b) &= \Pr[X \leq a, Y \leq b] \\ &= \Pr[\{\omega \in \Omega|\ (X(\omega) \leq a)\ \wedge\ (Y(\omega) \leq b)\}]\end{aligned}$$



Joint pdf can be obtained if they are jointly continuous

Equation:

$$\mathrm{F}_{X,,,Y}\ (a,b) = \int_{-\infty}^{b}\int_{-\infty}^{a} \mathrm{f}_{X,Y}\ (x,y)\ \mathrm{d}\ x\ \mathrm{d}\ y$$

(e.g., $\mathrm{f}_{X,Y}\ (x,y) = \frac{\partial^2 \mathrm{F}_{X,,,Y}(x,y)}{\partial x \partial y}$)

Joint pmf if they are jointly discrete

Equation:

$$\mathrm{p}_{X,Y}\ (x_k, y_l) = \Pr[X = x_k, Y = y_l]$$

Conditional density function

Equation:

$$f_{Y|X}(y|x) = \frac{\mathrm{f}_{X,Y}\ (x,y)}{\mathrm{f}_X\ (x)}$$

for all x with $\mathrm{f}_X\ (x) > 0$ otherwise conditional density is not defined for those values of x with $\mathrm{f}_X\ (x) = 0$

Two random variables are **independent** if
Equation:

$$\mathrm{f}_{X,Y}\ (x,y) = \mathrm{f}_X\ (x)\ \mathrm{f}_Y\ (y)$$

for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$. For discrete random variables,
Equation:

$$\mathrm{p}_{X,Y}\ (x_k, y_l) = \mathrm{p}_X\ (x_k)\ \mathrm{p}_Y\ (y_l)$$

for all k and l.

Moments

Statistical quantities to represent some of the characteristics of a random variable.
Equation:

$$\begin{aligned}\overline{g(X)} &= E[g(X)] \\ &= \begin{cases}\int_{-\infty}^{\infty} g(x)\ \mathrm{f}_X\ (x)\ \mathrm{d}\ x & \text{if continuous} \\ \sum_k g(x_k)\ \mathrm{p}_X\ (x_k) & \text{if discrete}\end{cases}\end{aligned}$$

- Mean
  Equation:

  $$\mu_X = \overline{X}$$

- Second moment
  Equation:

  $$E\left[X^2\right] = \overline{X^2}$$

- Variance
  Equation:

$$\begin{aligned}\operatorname{Var}(X) &= \mathrm{σ}(X)^2 \\ &= \overline{(X-\mu_X)^2} \\ &= \overline{X^2} - {\mu_X}^2\end{aligned}$$

- Characteristic function
  Equation:

  $$\mathit{\Phi}_X(u) = \overline{e^{iuX}}$$

  for $u \in \mathbb{R}$, where $i = \sqrt{-1}$
- Correlation between two random variables
  Equation:

  $$\begin{aligned}R_{XY} &= \overline{XY^*} \\ &= \begin{cases}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy^* \ \mathrm{f}_{X,Y} \ (x,y) \ \mathrm{d}\ x \ \mathrm{d}\ y & \text{if X and Y are jointly continuous} \\ \sum_k \sum_l x_k {y_l}^* \ \mathrm{p}_{X,Y} \ (x_k, y_l) & \text{if X and Y are jointly discrete}\end{cases}\end{aligned}$$

- Covariance
  Equation:

  $$\begin{aligned}C_{XY} &= \operatorname{Cov}(X,Y) \\ &= \overline{(X-\mu_X)(Y-\mu_Y)^*} \\ &= R_{XY} - \mu_X {\mu_Y}^*\end{aligned}$$

- Correlation coefficient
  Equation:

  $$\rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Uncorrelated random variables
  Two random variables X and Y are uncorrelated if $\rho_{XY} = 0$.

Introduction to Stochastic Processes

Definitions, distributions, and stationarity

Stochastic Process

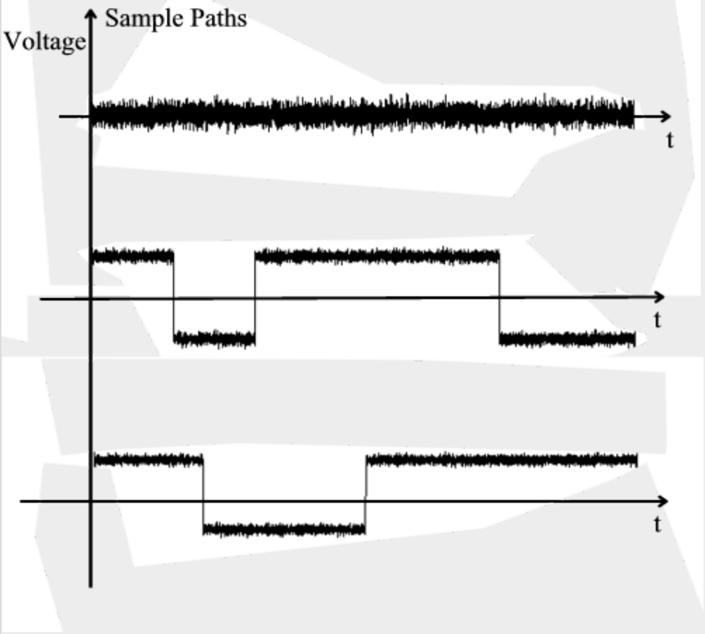
Given a sample space, a stochastic process is an indexed collection of random variables defined for each $\omega \in \Omega$.

Equation:

$$\forall t, t \in \mathbb{R} : (X_t(\omega))$$

Example:

Received signal at an antenna as in [link].



For a given t, $X_t(\omega)$ is a random variable with a distribution

Equation:

First-order distribution

$$\begin{aligned} F_{X_t}(b) &= \Pr[X_t \leq b] \\ &= \Pr[\{\omega \in \Omega | X_t(\omega) \leq b\}] \end{aligned}$$

First-order stationary process

If $F_{X_t}(b)$ is not a function of time then X_t is called a first-order stationary process.

Equation:

Second-order distribution

$$F_{X_{t_1}, X_{t_2}}(b_1, b_2) = \Pr[X_{t_1} \leq b_1, X_{t_2} \leq b_2]$$

for all $t_1 \in \mathbb{R}$, $t_2 \in \mathbb{R}$, $b_1 \in \mathbb{R}$, $b_2 \in \mathbb{R}$
Equation:

Nth-order distribution

$$F_{X_{t_1}, X_{t_2}, \ldots, X_{t_N}}(b_1, b_2, \ldots, b_N) = \mathrm{Pr}[X_{t_1} \leq b_1, \ldots, X_{t_N} \leq b_N]$$

Nth-order stationary : A random process is stationary of order N if
Equation:

$$F_{X_{t_1}, X_{t_2}, \ldots, X_{t_N}}(b_1, b_2, \ldots, b_N) = F_{X_{t_1+T}, X_{t_2+T}, \ldots, X_{t_N+T}}(b_1, b_2, \ldots, b_N)$$

Strictly stationary : A process is strictly stationary if it is Nth order stationary for all N.

Example:
$X_t = \cos(2\pi f_0 t + \Theta(\omega))$ where f_0 is the deterministic carrier frequency and $\Theta(\omega) : \Omega \to \mathbb{R}$ is a random variable defined over $[-\pi, \pi]$ and is assumed to be a uniform random variable; i.e.,
$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi} & \text{if } \theta \in [-\pi, \pi] \\ 0 & \text{otherwise} \end{cases}$
Equation:

$$\begin{aligned} F_{X_t}(b) &= \mathrm{Pr}[X_t \leq b] \\ &= \mathrm{Pr}[\cos(2\pi f_0 t + \Theta) \leq b] \end{aligned}$$

Equation:

$$F_{X_t}(b) = \mathrm{Pr}[-\pi \leq 2\pi f_0 t + \Theta \leq -\arccos(b)] + \mathrm{Pr}[\arccos(b) \leq 2\pi f_0 t + \Theta \leq \pi]$$

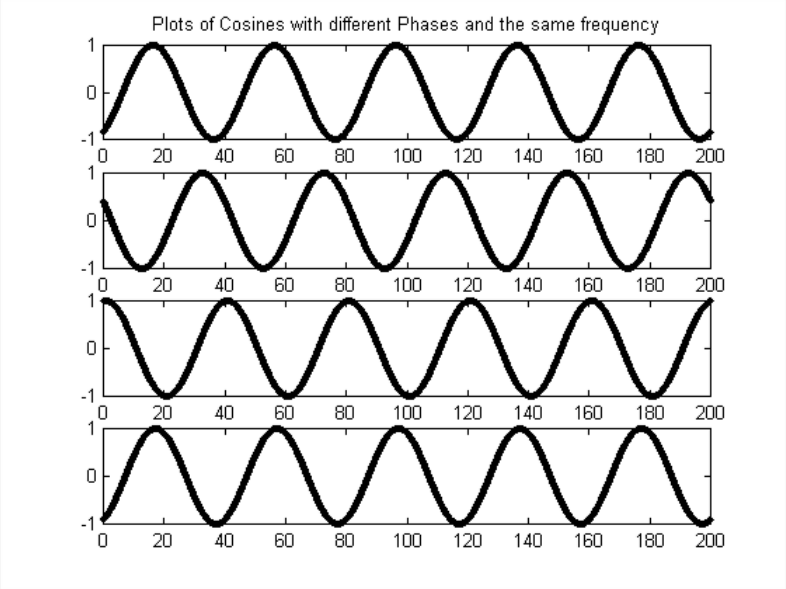
Equation:

$$\begin{aligned} F_{X_t}(b) &= \int_{(-\pi)-2\pi f_0 t}^{(-\arccos(b))-2\pi f_0 t} \frac{1}{2\pi}\,\mathrm{d}\,\theta + \int_{\arccos(b)-2\pi f_0 t}^{\pi-2\pi f_0 t} \frac{1}{2\pi}\,\mathrm{d}\,\theta \\ &= (2\pi - 2\arccos(b))\frac{1}{2\pi} \end{aligned}$$

Equation:

$$\begin{aligned} f_{X_t}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(1 - \frac{1}{\pi}\arccos(x)\right) \\ &= \begin{cases} \frac{1}{\pi\sqrt{1-x^2}} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

This process is stationary of order 1.



The second order stationarity can be determined by first considering conditional densities and the joint density. Recall that

Equation:

$$X_t = \cos(2\pi f_0 t + \Theta)$$

Then the relevant step is to find

Equation:

$$\Pr[X_{t_2} \leq b_2 \mid X_{t_1} = x_1]$$

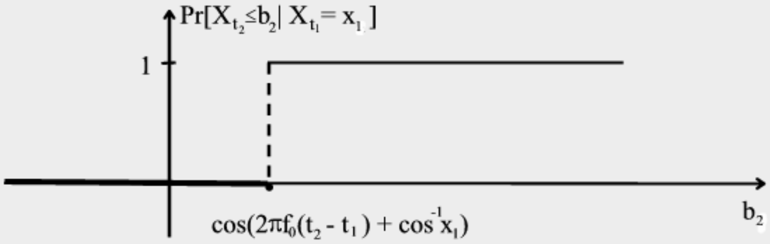
Note that

Equation:

$$(X_{t_1} = x_1 = \cos(2\pi f_0 t + \Theta)) \Rightarrow (\Theta = \arccos(x_1) - 2\pi f_0 t)$$

Equation:

$$\begin{aligned} X_{t_2} &= \cos(2\pi f_0 t_2 + \arccos(x_1) - 2\pi f_0 t_1) \\ &= \cos(2\pi f_0 \, (t_2 - t_1) + \arccos(x_1)) \end{aligned}$$

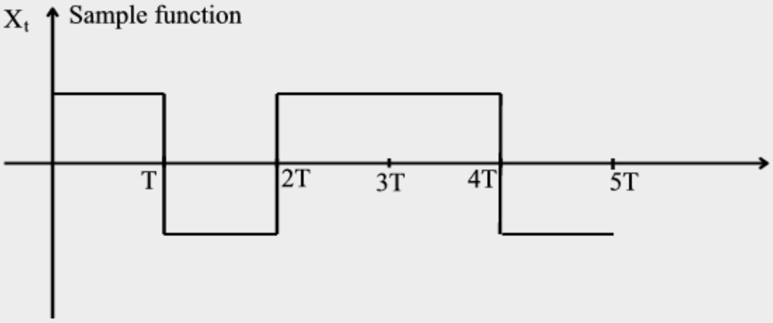


Equation:

$$F_{X_{t_2},X_{t_1}}(b_2,b_1) = \int_{-\infty}^{b_1} f_{X_{t_1}}(x_1)\,\mathrm{Pr}[X_{t_2} \leq b_2 \mid X_{t_1} = x_1]\;\mathrm{d}\,x_1$$

Note that this is only a function of $t_2 - t_1$.

Example:
Every T seconds, a fair coin is tossed. If heads, then $X_t = 1$ for $nT \leq t < (n+1)T$. If tails, then $X_t = -1$ for $nT \leq t < (n+1)T$.



Equation:

$$p_{X_t}(x) = \begin{cases} \frac{1}{2} & \text{if } \; x = 1 \\ \frac{1}{2} & \text{if } \; x = -1 \end{cases}$$

for all $t \in \mathbb{R}$. X_t is stationary of order 1.
Second order probability mass function
Equation:

$$p_{X_{t_1}X_{t_2}}(x_1,x_2) = p_{X_{t_2}|X_{t_1}}(x_2|x_1)p_{X_{t_1}}(x_1)$$

The conditional pmf
Equation:

$$p_{X_{t_2}|X_{t_1}}(x_2|x_1) = \begin{cases} 0 & \text{if } \; x_2 \neq x_1 \\ 1 & \text{if } \; x_2 = x_1 \end{cases}$$

when $nT \leq t_1 < (n+1)T$ and $nT \leq t_2 < (n+1)T$ for some n.
Equation:

$$p_{X_{t_2}|X_{t_1}}(x_2|x_1) = p_{X_{t_2}}(x_2)$$

for all x_1 and for all x_2 when $nT \leq t_1 < (n+1)T$ and $mT \leq t_2 < (m+1)T$ with $n \neq m$
Equation:

$$p_{X_{t_2}X_{t_1}}(x_2,x_1) = \begin{cases} 0 & \text{if } \; x_2 \neq x_1 \text{for } nT \leq t_1, t_2 < (n+1)T \\ p_{X_{t_1}}(x_1) & \text{if } \; x_2 = x_1 \text{for } nT \leq t_1, t_2 < (n+1)T \\ p_{X_{t_1}}(x_1)p_{X_{t_2}}(x_2) & \text{if } \; n \neq m \text{for } \; (nT \leq t_1 < (n+1)T) \; \wedge \; (mT \leq t_2 < (m+1)T) \end{cases}$$

Second-order Description

Second-order description

Practical and incomplete statistics

Mean

The mean function of a random process X_t is defined as the expected value of X_t for all t's.

Equation:

$$\begin{aligned} \mu_{X_t} &= E[X_t] \\ &= \begin{cases} \int_{-\infty}^{\infty} x \, \mathrm{f}_{X_t}\ (x) \, \mathrm{d}\ x \ \text{ if } \ \text{continuous} \\ \sum_{k=-\infty}^{\infty} x_k \, \mathrm{p}_{X_t}\ (x_k) \ \text{ if } \ \text{discrete} \end{cases} \end{aligned}$$

Autocorrelation

The autocorrelation function of the random process X_t is defined as

Equation:

$$\begin{aligned} R_X(t_2, t_1) &= E\left[X_{t_2}\overline{X_{t_1}}\right] \\ &= \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 \overline{x_1} \, \mathrm{f}_{X_{t_2}, X_{t_1}}\ (x_2, x_1) \, \mathrm{d}\ x_1 \, \mathrm{d}\ x_2 \ \text{ if } \ \text{continuous} \\ \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} x_l \overline{x_k} \, \mathrm{p}_{X_{t_2}, X_{t_1}}\ (x_l, x_k) \ \text{ if } \ \text{discrete} \end{cases} \end{aligned}$$

Fact

If X_t is second-order stationary, then $R_X(t_2, t_1)$ only depends on $t_2 - t_1$.

Equation:

$$\begin{aligned} R_X(t_2, t_1) &= E\left[X_{t_2}\overline{X_{t_1}}\right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 \overline{x_1} \, \mathrm{f}_{X_{t_2}, X_{t_1}}\ (x_2, x_1) \, \mathrm{d}\ x_2 \, \mathrm{d}\ x_1 \end{aligned}$$

Equation:

$$\begin{aligned} R_X(t_2, t_1) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 \overline{x_1} \, \mathrm{f}_{X_{t_2 - t_1}, X_0} \, (x_2, x_1) \, \mathrm{d} \, x_2 \, \mathrm{d} \, x_1 \\ &= R_X(t_2 - t_1, 0) \end{aligned}$$

If $R_X(t_2, t_1)$ depends on $t_2 - t_1$ only, then we will represent the autocorrelation with only one variable $\tau = t_2 - t_1$

Equation:

$$\begin{aligned} R_X(\tau) &= R_X(t_2 - t_1) \\ &= R_X(t_2, t_1) \end{aligned}$$

Properties

1. $R_X(0) \geq 0$
2. $R_X(\tau) = \overline{R_X(-\tau)}$
3. $|R_X(\tau)| \leq R_X(0)$

Example:

$X_t = \cos(2\pi f_0 t + \mathit{\Theta}(\omega))$ and $\mathit{\Theta}$ is uniformly distributed between 0 and 2π. The mean function

Equation:

$$\begin{aligned} \mu_X(t) &= E[X_t] \\ &= E[\cos(2\pi f_0 t + \mathit{\Theta})] \\ &= \int_0^{2\pi} \cos(2\pi f_0 t + \theta) \frac{1}{2\pi} \, \mathrm{d} \, \theta \\ &= 0 \end{aligned}$$

The autocorrelation function

Equation:

$$\begin{aligned} R_X(t + \tau, t) &= E\left[X_{t+\tau} \overline{X_t}\right] \\ &= E[\cos(2\pi f_0 \, (t + \tau) + \mathit{\Theta}) \cos(2\pi f_0 t + \mathit{\Theta})] \\ &= 1/2 E[\cos(2\pi f_0 \tau)] + 1/2 E[\cos(2\pi f_0 \, (2t + \tau) + 2\mathit{\Theta})] \\ &= 1/2 \cos(2\pi f_0 \tau) + 1/2 \int_0^{2\pi} \cos(2\pi f_0 \, (2t + \tau) + 2\theta) \frac{1}{2\pi} \, \mathrm{d} \, \theta \\ &= 1/2 \cos(2\pi f_0 \tau) \end{aligned}$$

Not a function of t since the second term in the right hand side of the equality in [link] is zero.

Example:
Toss a fair coin every T seconds. Since X_t is a discrete valued random process, the statistical characteristics can be captured by the pmf and the mean function is written as

Equation:

$$\begin{aligned}\mu_X(t) &= E[X_t] \\ &= 1/2 \times -1 + 1/2 \times 1 \\ &= 0\end{aligned}$$

Equation:

$$\begin{aligned}R_X(t_2, t_1) &= \sum_{kk} \sum_{ll} x_k x_l \; \mathrm{p}_{X_{t_2}, X_{t_1}} \; (x_k, x_l) \\ &= 1 \times 1 \times 1/2 - 1 \times -1 \times 1/2 \\ &= 1\end{aligned}$$

when $nT \le t_1 < (n+1)T$ and $nT \le t_2 < (n+1)T$

Equation:

$$\begin{aligned}R_X(t_2, t_1) &= 1 \times 1 \times 1/4 - 1 \times -1 \times 1/4 - 1 \times 1 \times 1/4 + 1 \times -1 \times 1/4 \\ &= 0\end{aligned}$$

when $nT \le t_1 < (n+1)T$ and $mT \le t_2 < (m+1)T$ with $n \ne m$

Equation:

$$R_X(t_2, t_1) = \begin{cases} 1 \;\text{ if }\; (nT \le t_1 < (n+1)T) \;\wedge\; (nT \le t_2 < (n+1)T) \\ 0 \;\text{ otherwise} \end{cases}$$

A function of t_1 and t_2.

Wide Sense Stationary

A process is said to be wide sense stationary if μ_X is constant and $R_X(t_2, t_1)$ is only a function of $t_2 - t_1$.

Fact

If X_t is strictly stationary, then it is wide sense stationary. The converse is not necessarily true.

Autocovariance

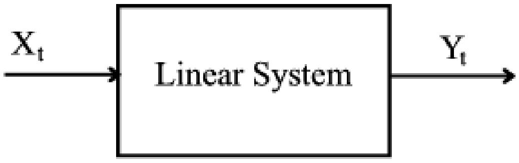
Autocovariance of a random process is defined as

Equation:

$$\begin{aligned} C_X(t_2, t_1) &= E\left[(X_{t_2} - \mu_X(t_2))\overline{X_{t_1} - \mu_X(t_1)}\right] \\ &= R_X(t_2, t_1) - \mu_X(t_2)\overline{\mu_X(t_1)} \end{aligned}$$

The variance of X_t is $\mathrm{Var}\,(X_t) = C_X(t, t)$

Two processes defined on one experiment ([link]).



Crosscorrelation

The crosscorrelation function of a pair of random processes is defined as

Equation:

$$\begin{aligned} R_{XY}(t_2, t_1) &= E\left[X_{t_2}\overline{Y_{t_1}}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, \mathrm{f}_{X_{t_2}, Y_{t_1}}\, (x, y)\ \mathrm{d}\, x\ \mathrm{d}\, y \end{aligned}$$

Equation:

$$C_{XY}(t_2, t_1) = R_{XY}(t_2, t_1) - \mu_X(t_2)\overline{\mu_Y(t_1)}$$

Jointly Wide Sense Stationary

The random processes X_t and Y_t are said to be jointly wide sense stationary if $R_{XY}(t_2, t_1)$ is a function of $t_2 - t_1$ only and $\mu_X(t)$ and $\mu_Y(t)$ are constant.

Gaussian Random Processes

Gaussian process

A process with mean $\mu_X(t)$ and covariance function $C_X(t_2, t_1)$ is said to be a Gaussian process if **any** $\boldsymbol{X} = (X_{t_1} X_{t_2} \ldots X_{t_N})^T$ formed by **any** sampling of the process is a Gaussian random vector, that is, **Equation:**

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} (\det \mathit{\Sigma}_X)^{\frac{1}{2}}} e^{-\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_X)^T \mathit{\Sigma}_X{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_X)\right)}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ where

$$\boldsymbol{\mu}_X = \begin{matrix} \mu_X(t_1) \\ \vdots \\ \mu_X(t_N) \end{matrix}$$

and

$$\mathit{\Sigma}_X = \begin{matrix} C_X(t_1, t_1) & \ldots & C_X(t_1, t_N) \\ \vdots & \ddots & \\ C_X(t_N, t_1) & \ldots & C_X(t_N, t_N) \end{matrix}$$

. The complete statistical properties of X_t can be obtained from the second-order statistics.

Properties

1. If a Gaussian process is WSS, then it is strictly stationary.
2. If two Gaussian processes are uncorrelated, then they are also statistically independent.
3. Any linear processing of a Gaussian process results in a Gaussian process.

Example:
X and Y are Gaussian and zero mean and independent. $Z = X + Y$ is also Gaussian.
Equation:

$$\begin{aligned}\varphi_X(u) &= e^{iuX} \\ &= e^{-\left(\frac{u^2}{2}\sigma_X^2\right)}\end{aligned}$$

for all $u \in \mathbb{R}$
Equation:

$$\begin{aligned}\varphi_Z(u) &= e^{iu(X+Y)} \\ &= e^{-\left(\frac{u^2}{2}\sigma_X^2\right)} e^{-\left(\frac{u^2}{2}\sigma_Y^2\right)} \\ &= e^{-\left(\frac{u^2}{2}\left(\sigma_X^2+\sigma_Y^2\right)\right)}\end{aligned}$$

therefore Z is also Gaussian.

White and Coloured Processes

White Noise

If we have a zero-mean Wide Sense Stationary process X, it is a **White Noise Process** if its ACF is a delta function at $\tau = 0$, i.e. it is of the form:
Equation:

$$r_{XX}(\tau) = P_X \delta(\tau)$$

where P_X is a constant.

The PSD of X is then given by
Equation:

$$\begin{aligned} S_X(\omega) &= \int P_X \delta(\tau) e^{-(i\omega\tau)} \, \mathrm{d}\,\tau \\ &= P_X e^{-(i\omega 0)} \\ &= P_X \end{aligned}$$

Hence X is **white**, since it contains equal power at **all** frequencies, as in **white light**.

P_X is the PSD of X at all frequencies.

But:
Equation:

$$\begin{aligned} \text{Power of X} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) \, \mathrm{d}\,\omega \\ &= \infty \end{aligned}$$

so the White Noise Process is unrealizable in practice, because of its infinite bandwidth.

However, it is very useful as a conceptual entity and as an approximation to 'nearly white' processes which have finite bandwidth, but which are 'white' over all frequencies of practical interest. For 'nearly white' processes, $r_{XX}(\tau)$ is a narrow pulse of non-zero width, and $S_X(\omega)$ is flat from zero up to some relatively high cutoff frequency and then decays to zero above that.

Strict Whiteness and i.i.d. Processes

Usually the above concept of whiteness is sufficient, but a much stronger definition is as follows:

Pick a set of times $\{t_1, t_2, \ldots, t_N\}$ to sample $X(t)$.

If, for **any choice** of $\{t_1, t_2, \ldots, t_N\}$ with N finite, the random variables $X(t_1)$, $X(t_2)$, $\ldots$ $X(t_N)$ are **jointly independent**, i.e. their joint pdf is given by
Equation:

$$f_{X(t_1),X(t_2),\ldots\,X(t_N)}(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f_{X(t_i)}(x_i)$$

and the marginal pdfs are identical, i.e.
Equation:

$$\begin{aligned} f_{X(t_1)} &= f_{X(t_2)} \\ &= \ldots \\ &= f_{X(t_N)} \\ &= f_X \end{aligned}$$

then the process is termed **Independent and Identically Distributed (i.i.d)**.

If, in addition, f_X is a pdf with zero mean, we have a **Strictly White Noise Process**.

An i.i.d. process is 'white' because the variables $X(t_i)$ and $X(t_j)$ are jointly independent, even when separated by an infinitesimally small interval between t_i and t_j.

Additive White Gaussian Noise (AWGN)

In many systems the concept of **Additive White Gaussian Noise (AWGN)** is used. This simply means a process which has a Gaussian pdf, a white PSD, and is linearly added to whatever signal we are analysing.

Note that although 'white' and Gaussian' often go together, this is **not necessary** (especially for 'nearly white' processes).

E.g. a very high speed random bit stream has an ACF which is approximately a delta function, and hence is a nearly white process, but its pdf is clearly not Gaussian - it is a pair of delta functions at $+(V)$ and $-V$, the two voltage levels of the bit stream.

Conversely a nearly white Gaussian process which has been passed through a lowpass filter (see next section) will still have a Gaussian pdf (as it is a summation of Gaussians) but will no longer be white.

Coloured Processes

A random process whose PSD is not white or nearly white, is often known as a **coloured noise** process.

We may obtain coloured noise $Y(t)$ with PSD $S_Y(\omega)$ simply by passing white (or nearly white) noise $X(t)$ with PSD P_X through a filter with frequency response $\mathscr{H}(\omega)$, such that from this equation from our discussion of Spectral Properties of Random Signals.
Equation:

$$\begin{aligned} S_Y(\omega) &= S_X(\omega)(|\mathscr{H}(\omega)|)^2 \\ &= P_X(|\mathscr{H}(\omega)|)^2 \end{aligned}$$

Hence if we design the filter such that
Equation:

$$|\mathscr{H}(\omega)| = \sqrt{\frac{S_Y(\omega)}{P_X}}$$

then $Y(t)$ will have the required coloured PSD.

For this to work, $S_Y(\omega)$ need only be constant (white) over the passband of the filter, so a **nearly white** process which satisfies this criterion is quite satisfactory and realizable.

Using this equation from our discussion of Spectral Properties of Random Signals and [link], the ACF of the coloured noise is given by
Equation:

$$\begin{aligned} r_{YY}(\tau) &= r_{XX}(\tau)^* h(-\tau)^* h(\tau) \\ &= P_X \delta(\tau)^* h(-\tau)^* h(\tau) \\ &= P_X h(-\tau)^* h(\tau) \end{aligned}$$

where $h(\tau)$ is the impulse response of the filter.

This Figure from previous discussion shows two examples of coloured noise, although the upper waveform is more 'nearly white' than the lower one, as can be seen in part c of this figure from previous discussion in which the upper PSD is flatter than the lower PSD. In these cases, the coloured waveforms were produced by passing uncorrelated random noise samples (white up to half the sampling frequency) through half-sine filters (as in this equation from our discussion of Random Signals) of length $T_b = 10$ and 50 samples respectively.

Linear Filtering

Equation:

Integration

$$Z(\omega) = \int_a^b X_t(\omega)\ \mathrm{d}\ t$$

Equation:

Linear Processing

$$Y_t = \int_{-\infty}^{\infty} h(t,\tau) X_\tau\ \mathrm{d}\ \tau$$

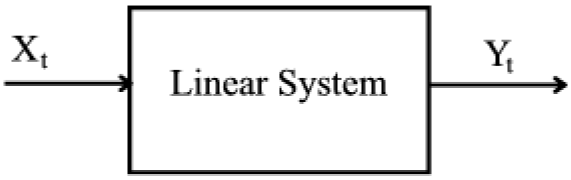
Equation:

Differentiation

$$X_t{}' = \frac{\mathrm{d}}{\mathrm{d}\ t}(X_t)$$

Properties

1. $Z = \int_a^b X_t(\omega)\ \mathrm{d}\ t = \int_a^b \mu_X(t)\ \mathrm{d}\ t$
2. $Z^2 = \int_a^b X_{t_2}\ \mathrm{d}\ t_2 \int_a^b X_{t_1}\ \mathrm{d}\ t_1 = \int_a^b \int_a^b R_X(t_2, t_1)\ \mathrm{d}\ t_1\ \mathrm{d}\ t_2$



Equation:

$$\begin{array}{rcl} \mu_Y(t) & = & \int_{-\infty}^{\infty} h(t,\tau) X_\tau \mathrm{~d~} \tau \\ & = & \int_{-\infty}^{\infty} h(t,\tau) \mu_X(\tau) \mathrm{~d~} \tau \end{array}$$

If X_t is wide sense stationary and the linear system is time invariant

Equation:

$$\begin{array}{rcl} \mu_Y(t) & = & \int_{-\infty}^{\infty} h(t-\tau) \mu_X \mathrm{~d~} \tau \\ & = & \mu_X \int_{-\infty}^{\infty} h(t') \mathrm{~d~} t' \\ & = & \mu_Y \end{array}$$

Equation:

$$\begin{array}{rcl} R_{YX}(t_2,t_1) & = & Y_{t_2} X_{t_1} \\ & = & \int_{-\infty}^{\infty} h(t_2-\tau) X_\tau \mathrm{~d~} \tau X_{t_1} \\ & = & \int_{-\infty}^{\infty} h(t_2-\tau) R_X(\tau-t_1) \mathrm{~d~} \tau \end{array}$$

Equation:

$$\begin{array}{rcl} R_{YX}(t_2,t_1) & = & \int_{-\infty}^{\infty} h(t_2-t_1-\tau') R_X(\tau') \mathrm{~d~} \tau' \\ & = & h{*}R_X(t_2-t_1) \end{array}$$

where $\tau' = \tau - t_1$.

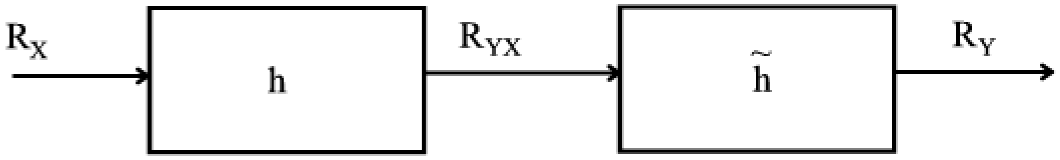
Equation:

$$\begin{array}{rcl} R_Y(t_2,t_1) & = & Y_{t_2} Y_{t_1} \\ & = & Y_{t_2} \int_{-\infty}^{\infty} h(t_1,\tau) X_\tau \mathrm{~d~} \tau \\ & = & \int_{-\infty}^{\infty} h(t_1,\tau) R_{YX}(t_2,\tau) \mathrm{~d~} \tau \\ & = & \int_{-\infty}^{\infty} h(t_1-\tau) R_{YX}(t_2-\tau) \mathrm{~d~} \tau \end{array}$$

Equation:

$$\begin{aligned} R_Y(t_2,t_1) &= \int_{-\infty}^{\infty} h(\tau' - (t_2 - t_1)) R_{YX}(\tau')\,\mathrm{d}\,\tau' \\ &= R_Y(t_2 - t_1) \\ &= \widetilde{h} * R_{YX}(t_2,t_1) \end{aligned}$$

where $\tau' = t_2 - \tau$ and $\widetilde{h}(\tau) = h(-\tau)$ for all $\tau \in \mathbb{R}$. Y_t is WSS if X_t is WSS and the linear system is time-invariant.



Example:
X_t is a wide sense stationary process with $\mu_X = 0$, and $R_X(\tau) = \frac{N_0}{2}\delta(\tau)$. Consider the random process going through a filter with impulse response $h(t) = e^{-(at)}u(t)$. The output process is denoted by Y_t. $\mu_Y(t) = 0$ for all t.

Equation:

$$\begin{aligned} R_Y(\tau) &= \frac{N_0}{2} \int_{-\infty}^{\infty} h(\alpha) h(\alpha - \tau)\,\mathrm{d}\,\alpha \\ &= \frac{N_0}{2} \frac{e^{-(a|\tau|)}}{2a} \end{aligned}$$

X_t is called a white process. Y_t is a Markov process.

Power Spectral Density

The power spectral density function of a wide sense stationary (WSS) process X_t is defined to be the Fourier transform of the autocorrelation function of X_t.
Equation:

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-(i2\pi f\tau)} \mathrm{~d}\, \tau$$

if X_t is WSS with autocorrelation function $R_X(\tau)$.

Properties

1. $S_X(f) = S_X(-f)$ since R_X is even and real.
2. $\text{Var}\,(X_t) = R_X(0) = \int_{-\infty}^{\infty} S_X(f) \mathrm{~d}\, f$
3. $S_X(f)$ is real and nonnegative $S_X(f) \geq 0$ for all f.

If $Y_t = \int_{-\infty}^{\infty} h(t-\tau) X_\tau \mathrm{~d}\, \tau$ then
Equation:

$$\begin{aligned} S_Y(f) &= \mathscr{F}(R_Y(\tau)) \\ &= \mathscr{F}\left(h^* \widetilde{h}^* R_X(\tau)\right) \\ &= H(f)\widetilde{H}(f) S_X(f) \\ &= (|H(f)|)^2 S_X(f) \end{aligned}$$

since $\widetilde{H}(f) = \int_{-\infty}^{\infty} \widetilde{h}(t) e^{-(i2\pi f t)} \mathrm{~d}\, t = H(f)$

Example:
X_t is a white process and $h(t) = e^{-(at)} u(t)$.
Equation:

$$H(f) = \frac{1}{a + i2\pi f}$$

Equation:

$$S_Y(f) = \frac{\frac{N_0}{2}}{a^2 + 4\pi^2 f^2}$$

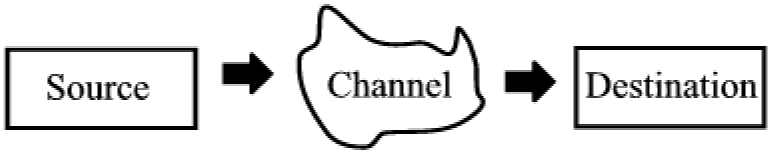
Information Theory and Coding

In the previous chapters, we considered the problem of digital transmission over different channels. Information sources are not often digital, and in fact, many sources are analog. Although many channels are also analog, it is still more efficient to convert analog sources into digital data and transmit over analog channels using digital transmission techniques. There are two reasons why digital transmission could be more efficient and more reliable than analog transmission:

1. Analog sources could be compressed to digital form efficiently.
2. Digital data can be transmitted over noisy channels reliably.

There are several key questions that need to be addressed:

1. How can one model information?
2. How can one quantify information?
3. If information can be measured, does its information quantity relate to how much it can be compressed?
4. Is it possible to determine if a particular channel can handle transmission of a source with a particular information quantity?



Example:
The information content of the following sentences: "Hello, hello, hello." and "There is an exam today." are not the same. Clearly the second one carries more information. The first one can be compressed to "Hello" without much loss of information.

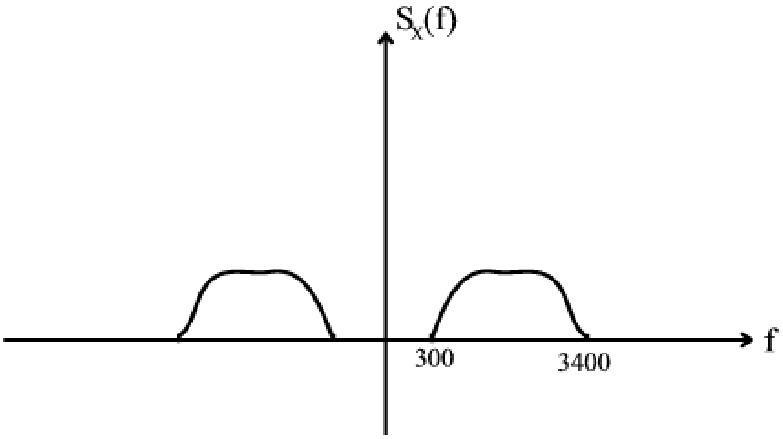
In other modules, we will quantify information and find efficient representation of information (Entropy). We will also quantify how much information can be transmitted through channels, reliably. Channel coding can be used to reduce information rate and increase reliability.

Entropy

Information sources take very different forms. Since the information is not known to the destination, it is then best modeled as a random process, discrete-time or continuous time.

Here are a few examples:

- Digital data source (e.g., a text) can be modeled as a discrete-time and discrete valued random process X_1, X_2, …, where $X_i \in \{A, B, C, D, E, \ldots\}$ with a particular $p_{X_1}(x)$, $p_{X_2}(x)$, …, and a specific $p_{X_1 X_2}$, $p_{X_2 X_3}$, …, and $p_{X_1 X_2 X_3}$, $p_{X_2 X_3 X_4}$, …, etc.
- Video signals can be modeled as a continuous time random process. The power spectral density is bandlimited to around 5 MHz (the value depends on the standards used to raster the frames of image).
- Audio signals can be modeled as a continuous-time random process. It has been demonstrated that the power spectral density of speech signals is bandlimited between 300 Hz and 3400 Hz. For example, the speech signal can be modeled as a Gaussian process with the shown power spectral density over a small observation period.



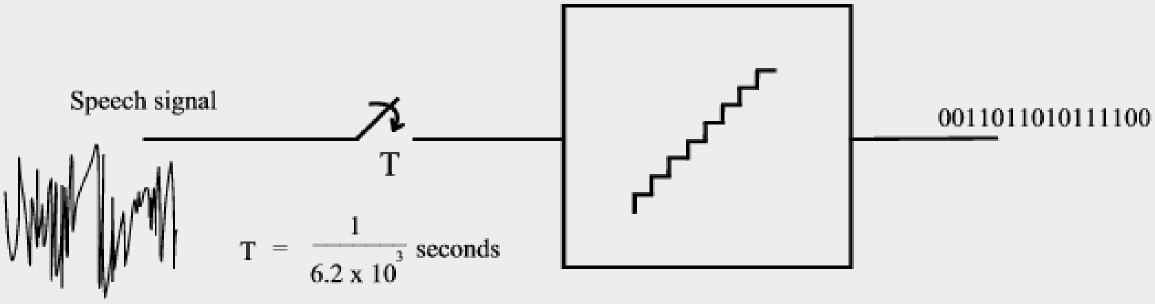
These analog information signals are bandlimited. Therefore, if sampled faster than the Nyquist rate, they can be reconstructed from their sample values.

Example:
A speech signal with bandwidth of 3100 Hz can be sampled at the rate of 6.2 kHz. If the samples are quantized with a 8 level quantizer then the speech signal can be

represented with a binary sequence with the rate of

Equation:

$$\begin{aligned} 6.2 \times 10^3 \log_2 8 &= 18600 \frac{\text{bits}}{\text{sample}} \frac{\text{samples}}{\text{sec}} \\ &= 18.6 \frac{\text{kbits}}{\text{sec}} \end{aligned}$$



The sampled real values can be quantized to create a discrete-time discrete-valued random process. Since any bandlimited analog information signal can be converted to a sequence of discrete random variables, we will continue the discussion only for discrete random variables.

Example:

The random variable $\boldsymbol{x}$ takes the value of 0 with probability 0.9 and the value of 1 with probability 0.1. The statement that $\boldsymbol{x} = 1$ carries more information than the statement that $\boldsymbol{x} = 0$. The reason is that $\boldsymbol{x}$ is expected to be 0, therefore, knowing that $\boldsymbol{x} = 1$ is more surprising news!! An intuitive definition of information measure should be larger when the probability is small.

Example:

The information content in the statement about the temperature and pollution level on July 15th in Chicago should be the sum of the information that July 15th in Chicago was hot and highly polluted since pollution and temperature could be independent.

Equation:

$$I(\text{hot}, \text{high}) = I(\text{hot}) + I(\text{high})$$

An intuitive and meaningful measure of information should have the following properties:

1. Self information should decrease with increasing probability.
2. Self information of two independent events should be their sum.
3. Self information should be a continuous function of the probability.

The only function satisfying the above conditions is the -log of the probability.

Entropy

The entropy (average self information) of a discrete random variable X is a function of its probability mass function and is defined as

Equation:

$$H(X) = -\sum_{i=1}^{N} \mathrm{p}_X\ (x_i) \log \mathrm{p}_X\ (x_i)$$

where N is the number of possible values of X and $\mathrm{p}_X\ (x_i) = \Pr[X = x_i]$. If log is base 2 then the unit of entropy is bits. Entropy is a measure of uncertainty in a random variable and a measure of information it can reveal. A more basic explanation of entropy is provided in another module.

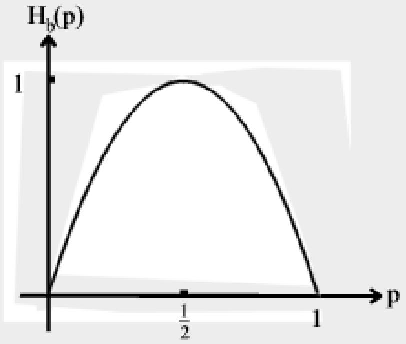
Example:

If a source produces binary information $\{0, 1\}$ with probabilities p and $1 - p$. The entropy of the source is

Equation:

$$H(X) = (-(p \log_2 p)) - (1 - p) \log_2 (1 - p)$$

If $p = 0$ then $H(X) = 0$, if $p = 1$ then $H(X) = 0$, if $p = 1/2$ then $H(X) = 1$ bits. The source has its largest entropy if $p = 1/2$ and the source provides no new information if $p = 0$ or $p = 1$.



Example:
An analog source is modeled as a continuous-time random process with power spectral density bandlimited to the band between 0 and 4000 Hz. The signal is sampled at the Nyquist rate. The sequence of random variables, as a result of sampling, are assumed to be independent. The samples are quantized to 5 levels $\{-2, -1, 0, 1, 2\}$. The probability of the samples taking the quantized values are $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right\}$, respectively. The entropy of the random variables are
Equation:

$$\begin{aligned} H(X) &= \left(-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right) - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{1}{16}\log_2\frac{1}{16} \\ &= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{16}\log_2 16 + \frac{1}{16}\log_2 16 \\ &= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{4}{8} \\ &= \frac{15}{8}\frac{\text{bits}}{\text{sample}} \end{aligned}$$

There are 8000 samples per second. Therefore, the source produces $8000 \times \frac{15}{8} = 15000\frac{\text{bits}}{\text{sec}}$ of information.

Joint Entropy

The joint entropy of two discrete random variables (X, Y) is defined by
Equation:

$$H(X,Y) = -\sum_{ii}\sum_{jj} \mathrm{p}_{X,Y}\ (x_i, y_j)\log \mathrm{p}_{X,Y}\ (x_i, y_j)$$

The joint entropy for a random vector $\boldsymbol{X} = (X_1 X_2 ... X_n)^T$ is defined as

Equation:

$$H(\boldsymbol{X}) = -\sum_{x_1 x_1}\sum_{x_2 x_2}\cdots\sum_{x_n x_n} \mathrm{p}_{\boldsymbol{X}}\ (x_1, x_2, ..., x_n) \log\, \mathrm{p}_{\boldsymbol{X}}\ (x_1, x_2, ..., x_n)$$

Conditional Entropy

The conditional entropy of the random variable X given the random variable Y is defined by

Equation:

$$H(X|Y) = -\sum_{ii}\sum_{jj} \mathrm{p}_{X,Y}\ (x_i, y_j) \log p_{X|Y}(x_i|y_j)$$

It is easy to show that

Equation:

$$H(\boldsymbol{X}) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1 X_2 ... X_{\text{n-1}})$$

and

Equation:

$$\begin{aligned} H(X,Y) &= H(Y) + H(X|Y) \\ &= H(X) + H(Y|X) \end{aligned}$$

If X_1, X_2, ..., X_n are mutually independent it is easy to show that

Equation:

$$H(\boldsymbol{X}) = \sum_{i=1}^{n} H(X_i)$$

Entropy Rate

The entropy rate of a stationary discrete-time random process is defined by

Equation:

$$H = \underset{n\to\infty}{\text{limit}}\ H(X_n|X_1 X_2 ... X_n)$$

The limit exists and is equal to
Equation:

$$H = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

The entropy rate is a measure of the uncertainty of information content per output symbol of the source.

Entropy is closely tied to source coding. The extent to which a source can be compressed is related to its entropy. In 1948, Claude E. Shannon introduced a theorem which related the entropy to the number of bits per second required to represent a source without much loss.

Source Coding

As mentioned earlier, how much a source can be compressed should be related to its entropy. In 1948, Claude E. Shannon introduced three theorems and developed very rigorous mathematics for digital communications. In one of the three theorems, Shannon relates entropy to the minimum number of bits per second required to represent a source without much loss (or distortion).

Consider a source that is modeled by a discrete-time and discrete-valued random process X_1, X_2, …, X_n, … where $x_i \in \{a_1, a_2, \ldots, a_N\}$ and define $p_{X_i}(x_i = a_j) = p_j$ for $j = 1$,2,. . .,N, where it is assumed that X_1, X_2,… X_n are mutually independent and identically distributed.

Consider a sequence of length n

Equation:

$$\boldsymbol{X} = \begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{matrix}$$

The symbol a_1 can occur with probability p_1. Therefore, in a sequence of length n, on the average, a_1 will appear np_1 times with high probabilities if n is very large.

Therefore,

Equation:

$$P(\boldsymbol{X} = \boldsymbol{x}) = p_{X_1}(x_1) p_{X_2}(x_2) \ldots p_{X_n}(x_n)$$

Equation:

$$P(\boldsymbol{X} = \boldsymbol{x}) \simeq {p_1}^{np_1} {p_2}^{np_2} \ldots {p_N}^{np_N} = \prod_{i=1}^{N} {p_i}^{np_i}$$

where $p_i = P(X_j = a_i)$ for all j and for all i.

A typical sequence $\boldsymbol{X}$ may look like
Equation:

$$\boldsymbol{X} = \begin{array}{c} a_2 \\ \vdots \\ a_1 \\ a_N \\ a_2 \\ a_5 \\ \vdots \\ a_1 \\ \vdots \\ a_N \\ a_6 \end{array}$$

where a_i appears np_i times with large probability. This is referred to as a **typical sequence**. The probability of $\boldsymbol{X}$ being a typical sequence is
Equation:

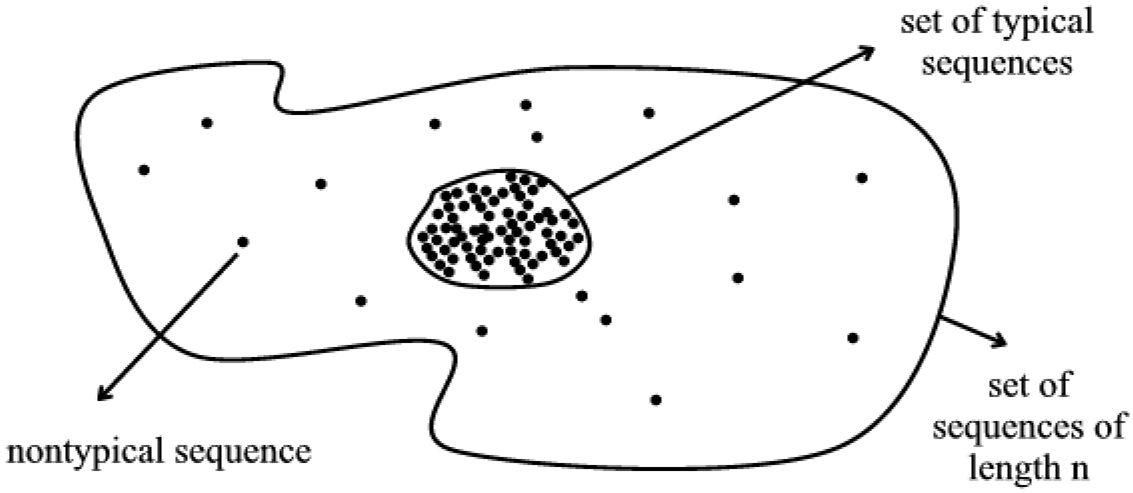
$$\begin{aligned} P(\boldsymbol{X} = \boldsymbol{x}) \simeq \prod_{i=1}^{N} {p_i}^{np_i} &= \prod_{i=1}^{N} \left(2^{\log_2 p_i}\right)^{np_i} \\ &= \prod_{i=1}^{N} 2^{np_i \log_2 p_i} \\ &= 2^{n \sum_{i=1}^{N} p_i \log_2 p_i} \\ &= 2^{-(nH(X))} \end{aligned}$$

where $H(X)$ is the entropy of the random variables X_1, X_2,…, X_n.

For large n, almost all the output sequences of length n of the source are equally probably with $\text{probability} \simeq 2^{-(nH(X))}$. These are typical sequences. The probability of nontypical sequences are negligible. There are N^n different sequences of length n with alphabet of size N. The probability of typical sequences is almost 1.

Equation:

$$\sum_{k=1}^{\#\text{ of typical seq.}} 2^{-(nH(X))} = 1$$



Example:

Consider a source with alphabet {A,B,C,D} with probabilities { $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$ }. Assume X_1, X_2,…, X_8 is an independent and identically distributed sequence with $X_i \in \{A, B, C, D\}$ with the above probabilities.

Equation:

$$\begin{aligned} H(X) &= \left(-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right)-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{8}\log_2\frac{1}{8} \\ &= \frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{3}{8} \\ &= \frac{4+4+6}{8} \\ &= \frac{14}{8} \end{aligned}$$

The number of typical sequences of length 8
Equation:

$$2^{8\times\frac{14}{8}}=2^{14}$$

The number of nontypical sequences
$4^8-2^{14}=2^{16}-2^{14}=2^{14}(4-1)=3\times2^{14}$
Examples of typical sequences include those with A appearing $8\times\frac{1}{2}=4$ times, B appearing $8\times\frac{1}{4}=2$ times, etc. {A,D,B,B,A,A,C,A}, {A,A,A,A,C,D,B,B} and much more.
Examples of nontypical sequences of length 8: {D,D,B,C,C,A,B,D}, {C,C,C,C,C,B,C,C} and much more. Indeed, these definitions and arguments are valid when n is very large. The probability of a source output to be in the set of typical sequences is 1 when $n\rightarrow\infty$. The probability of a source output to be in the set of nontypical sequences approaches 0 as $n\rightarrow\infty$.

The essence of source coding or data compression is that as $n\rightarrow\infty$, nontypical sequences never appear as the output of the source. Therefore, one only needs to be able to represent typical sequences as binary codes and ignore nontypical sequences. Since there are only $2^{nH(X)}$ typical sequences of length n, it takes $nH(X)$ bits to represent them on the average. On the average it takes $H(X)$ bits per source output to represent a simple source that produces independent and identically distributed outputs.
Theorem
Shannon's Source-Coding

A source that produced independent and identically distributed random variables with entropy H can be encoded with arbitrarily small error probability at any rate R in bits per source output if $R \geq H$. Conversely, if $R < H$, the error probability will be bounded away from zero, independent of the complexity of coder and decoder.

The source coding theorem proves existence of source coding techniques that achieve rates close to the entropy but does not provide any algorithms or ways to construct such codes.

If the source is not i.i.d. (independent and identically distributed), but it is stationary with memory, then a similar theorem applies with the entropy $H(X)$ replaced with the entropy rate $H = \underset{n \to \infty}{\text{limit}}\; H(X_n | X_1 X_2 \ldots X_{\text{n-1}})$

In the case of a source with memory, the more the source produces outputs the more one knows about the source and the more one can compress.

Example:
The English language has 26 letters, with space it becomes an alphabet of size 27. If modeled as a memoryless source (no dependency between letters in a word) then the entropy is $H(X) = 4.03$ bits/letter.
If the dependency between letters in a text is captured in a model the entropy rate can be derived to be $H = 1.3$ bits/letter. Note that a non-information theoretic representation of a text may require 5 bits/letter since 2^5 is the closest power of 2 to 27. Shannon's results indicate that there may be a compression algorithm with the rate of 1.3 bits/letter.

Although Shannon's results are not constructive, there are a number of source coding algorithms for discrete time discrete valued sources that come close to Shannon's bound. One such algorithm is the Huffman source coding algorithm. Another is the Lempel and Ziv algorithm.

Huffman codes and Lempel and Ziv apply to compression problems where the source produces discrete time and discrete valued outputs. For cases where the source is analog there are powerful compression algorithms that specify all the steps from sampling, quantizations, and binary representation. These are referred to as waveform coders. JPEG, MPEG, vocoders are a few examples for image, video, and voice, respectively.

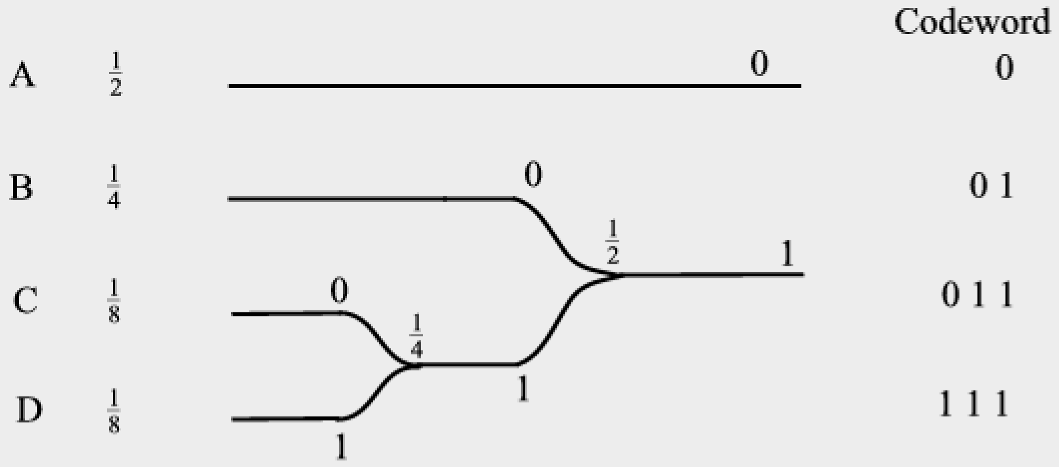
Huffman Coding

One particular source coding algorithm is the Huffman encoding algorithm. It is a source coding algorithm which approaches, and sometimes achieves, Shannon's bound for source compression. A brief discussion of the algorithm is also given in another module.

Huffman encoding algorithm

1. Sort source outputs in decreasing order of their probabilities
2. Merge the two least-probable outputs into a single output whose probability is the sum of the corresponding probabilities.
3. If the number of remaining outputs is more than 2, then go to step 1.
4. Arbitrarily assign 0 and 1 as codewords for the two remaining outputs.
5. If an output is the result of the merger of two outputs in a preceding step, append the current codeword with a 0 and a 1 to obtain the codeword the the preceding outputs and repeat step 5. If no output is preceded by another output in a preceding step, then stop.

Example:
X A B C D with probabilities { −,−,−,−}



— — — — —. As you may recall, the entropy of the source was also $H\ X$ —. In this case, the Huffman code achieves the lower bound of — ———.

In general, we can define average code length as

Equation:

$$ {}_{X}\;x\quad x $$
$$ {}_{x\;\;X} $$

where X is the set of possible values of x.

It is not very hard to show that

Equation:

$$ H\ X \qquad\qquad H\ X $$

For compressing single source output at a time, Huffman codes provide nearly optimum code lengths.

The drawbacks of Huffman coding

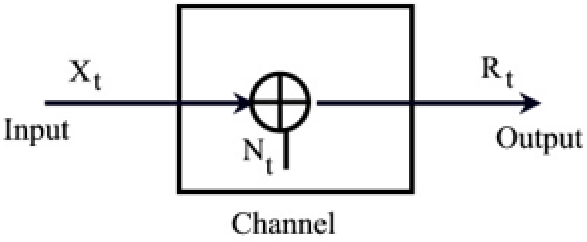
1. Codes are variable length.
2. The algorithm requires the knowledge of the probabilities, ${}_{X}\;x$ for all $x\quad X$.

Another powerful source coder that does not have the above shortcomings is Lempel and Ziv.

Data Transmission and Reception

We will develop the idea of **data transmission** by first considering simple channels. In additional modules, we will consider more practical channels; **baseband** channels with **bandwidth** constraints and **passband** channels.

Simple additive white Gaussian channels



carries data, is a white
Gaussian random process.

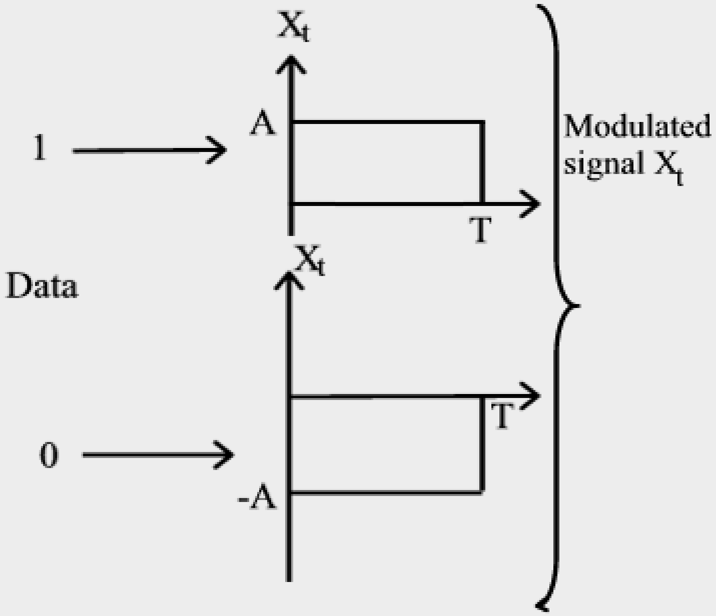
The concept of using different types of modulation for transmission of data is introduced in the module Signalling. The problem of demodulation and detection of signals is discussed in Demodulation and Detection.

Signalling

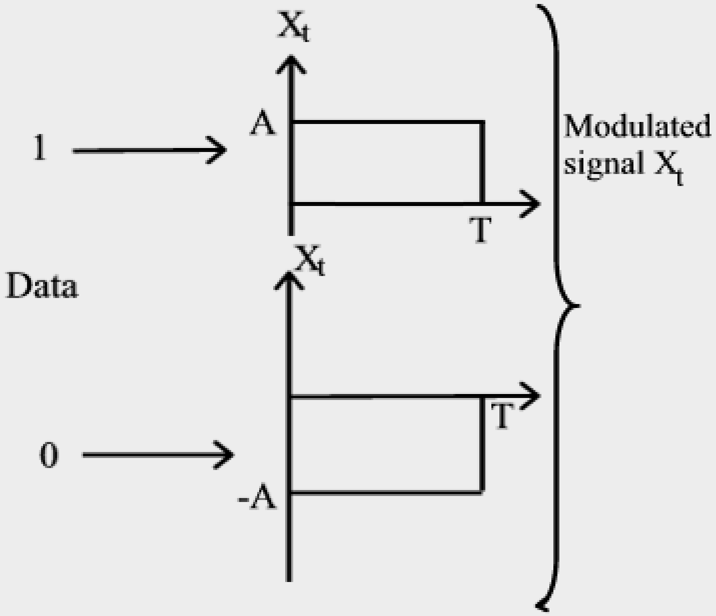
Example:

Data symbols are "1" or "0" and data rate is $\frac{1}{T}$ Hertz.

Pulse amplitude modulation (PAM)



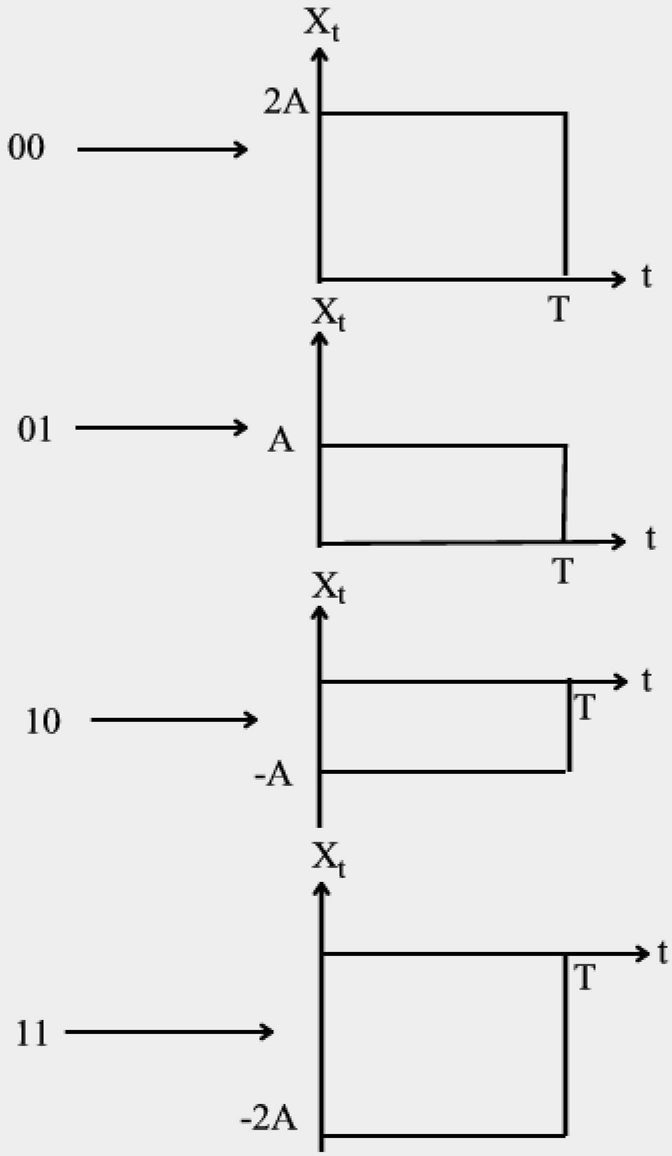
Pulse position modulation



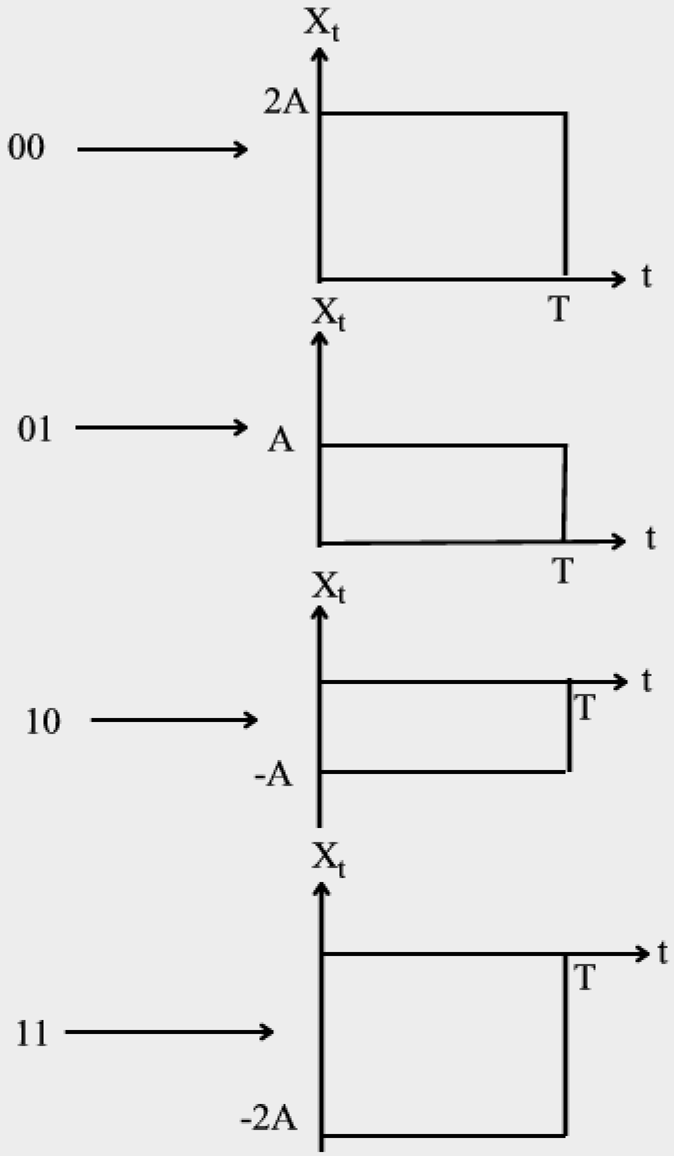
Example:

Example

Data symbols are "1" or "0" and the data rate is $\frac{2}{T}$ Hertz.



This strategy is an alternative to PAM with half the period, $\frac{T}{2}$.



Relevant measures are energy of modulated signals

Equation:

$$E_m = \forall m \in \{1, 2, \ldots, M\} : \left(\int_0^T {s_m}^2(t) \, \mathrm{d} \, t\right)$$

and how different they are in terms of inner products.

Equation:

$$\langle s_m, s_n \rangle = \int_0^T s_m(t) s_n(t)\,\mathrm{d}\,t$$

for $m \in \{1, 2, \ldots, M\}$ and $n \in \{1, 2, \ldots, M\}$.

antipodal

Signals $s_1(t)$ and $s_2(t)$ are antipodal if $\forall t, t \in [0, T] : (s_2(t) = -s_1(t))$

orthogonal

Signals $s_1(t)$, $s_2(t)$,…, $s_M(t)$ are orthogonal if $\langle s_m, s_n \rangle = 0$ for $m \neq n$.

biorthogonal

Signals $s_1(t)$, $s_2(t)$,…, $s_M(t)$ are biorthogonal if $s_1(t)$,…, $s_{\frac{M}{2}}(t)$ are orthogonal and $s_m(t) = -s_{\frac{M}{2}+m}(t)$ for some $m \in \left\{1, 2, \ldots, \frac{M}{2}\right\}$.

It is quite intuitive to expect that the smaller (the more negative) the inner products, $\langle s_m, s_n \rangle$ for all $m \neq n$, the better the signal set.

Simplex signals

Let $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ be a set of orthogonal signals with equal energy. The signals $\widetilde{s_1}(t)$,…, $\widetilde{s_M}(t)$ are simplex signals if

Equation:

$$\widetilde{s_m}(t) = s_m(t) - \frac{1}{M} \sum_{k=1}^{M} s_k(t)$$

If the energy of orthogonal signals is denoted by

Equation:

$$\forall m, m \in \{1, 2, ..., M\} : \left(E_s = \int_0^T {s_m}^2(t)\, \mathrm{d}\, t\right)$$

then the energy of simplex signals

Equation:

$$E_{\tilde{s}} = \left(1 - \frac{1}{M}\right) E_s$$

and

Equation:

$$\forall m \neq n : \left(\langle \widetilde{s_m}, \widetilde{s_n} \rangle = \frac{-1}{M-1} E_{\tilde{s}}\right)$$

It is conjectured that among all possible M-ary signals with equal energy, the simplex signal set results in the smallest probability of error when used to transmit information through an additive white Gaussian noise channel.

The geometric representation of signals can provide a compact description of signals and can simplify performance analysis of communication systems using the signals.

Once signals have been modulated, the receiver must detect and demodulate the signals despite interference and noise and decide which of the set of possible transmitted signals was sent.

Geometric Representation of Modulation Signals

Geometric representation of signals can provide a compact characterization of signals and can simplify analysis of their performance as modulation signals.

Orthonormal bases are essential in geometry. Let $\{s_1(t), s_2(t), ..., s_M(t)\}$ be a set of signals.

Define $\psi_1(t) = \frac{s_1(t)}{\sqrt{E_1}}$ where $E_1 = \int_0^T {s_1}^2(t) \text{ d } t$.

Define $s_{21} = \langle s_2, \psi_1 \rangle = \int_0^T s_2(t)\overline{\psi_1(t)} \text{ d } t$ and $\psi_2(t) = \frac{1}{\sqrt{\widehat{E_2}}} \left(s_2(t) - s_{21}\psi_1\right)$ where $\widehat{E_2} = \int_0^T \left(s_2(t) - s_{21}\psi_1(t)\right)^2 \text{ d } t$

In general
Equation:

$$\psi_k(t) = \frac{1}{\sqrt{\widehat{E_k}}} \left(s_k(t) - \sum_{j=1}^{k-1} s_{\text{kj}}\psi_j(t)\right)$$

where $\widehat{E_k} = \int_0^T \left(s_k(t) - \sum_{j=1}^{k-1} s_{\text{kj}}\psi_j(t)\right)^2 \text{ d } t$.

The process continues until all of the M signals are exhausted. The results are N orthogonal signals with unit energy, $\{\psi_1(t), \psi_2(t), ..., \psi_N(t)\}$ where $N \leq M$. If the signals $\{s_1(t), ..., s_M(t)\}$ are linearly independent, then $N = M$.

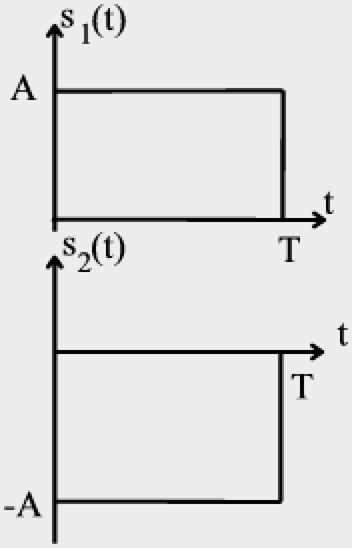
The M signals can be represented as
Equation:

$$s_m(t) = \sum_{n=1}^{N} s_{\text{mn}}\psi_n(t)$$

with $m \in \{1, 2, \ldots, M\}$ where $s_{\mathrm{mn}} = \langle s_m, \psi_n \rangle$ and $E_m = \sum_{n=1}^{N} {s_{\mathrm{mn}}}^2$.

The signals can be represented by $s_m = \begin{pmatrix} s_{m1} \\ s_{m2} \\ \vdots \\ s_{\mathrm{mN}} \end{pmatrix}$

Example:



Equation:

$$\psi_1(t) = \frac{s_1(t)}{\sqrt{A^2 T}}$$

Equation:

$$s_{11} = A\sqrt{T}$$

Equation:

$$s_{21} = -\left(A\sqrt{T}\right)$$

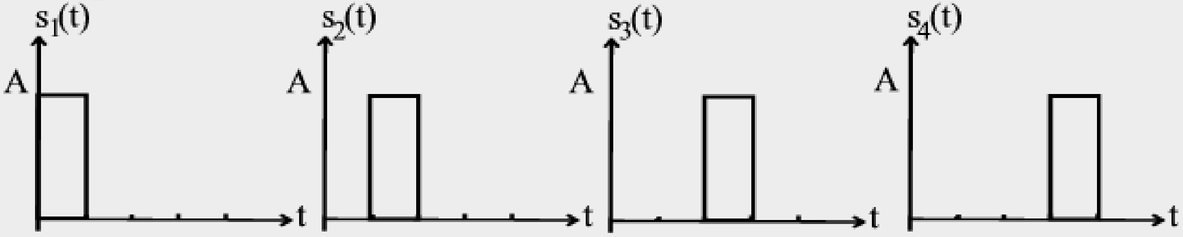
Equation:

$$\begin{aligned}\psi_2(t) &= (s_2(t) - s_{21}\psi_1(t))\frac{1}{\sqrt{\widehat{E_2}}} \\ &= \left(-A + \frac{A\sqrt{T}}{\sqrt{T}}\right)\frac{1}{\sqrt{\widehat{E_2}}} \\ &= 0\end{aligned}$$

X_t
Linear System
Y_t

Dimension of the signal set is 1 with $E_1 = {s_{11}}^2$ and $E_2 = {s_{21}}^2$.

Example:



$\psi_m(t) = \frac{s_m(t)}{\sqrt{E_s}}$ where $E_s = \int_0^T {s_m}^2(t)\,\mathrm{d}\,t = \frac{A^2T}{4}$

$s_1 = \begin{pmatrix}\sqrt{E_s}\\0\\0\\0\end{pmatrix}$, $s_2 = \begin{pmatrix}0\\\sqrt{E_s}\\0\\0\end{pmatrix}$, $s_3 = \begin{pmatrix}0\\0\\\sqrt{E_s}\\0\end{pmatrix}$, and $s_4 = \begin{pmatrix}0\\0\\0\\\sqrt{E_s}\end{pmatrix}$

Equation:

$$\forall mn : \left(d_{\mathrm{mn}} = |s_m - s_n| = \sqrt{\sum_{j=1}^{N} (s_{\mathrm{mj}} - s_{\mathrm{nj}})^2 = \sqrt{2E_s}}\right)$$

is the Euclidean distance between signals.

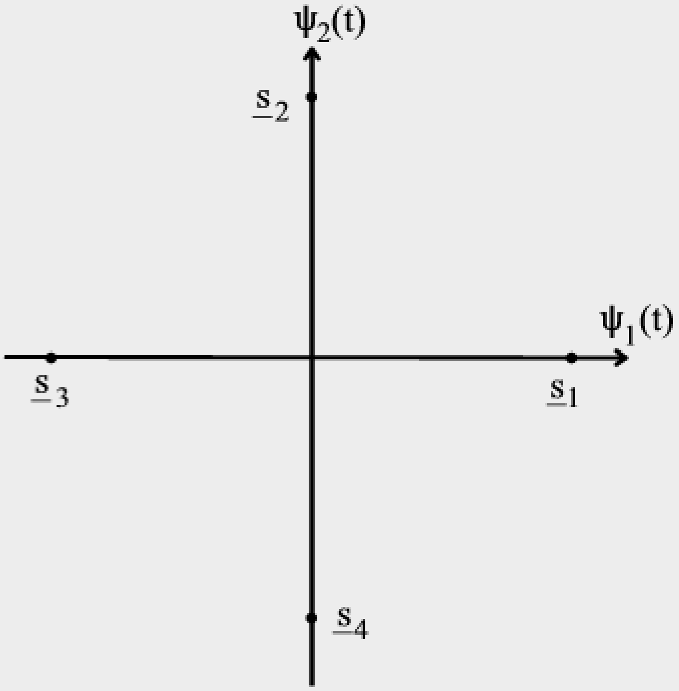
Example:

Set of 4 equal energy biorthogonal signals. $s_1(t) = s(t)$, $s_2(t) = s^{\perp}(t)$, $s_3(t) = -s(t)$, $s_4(t) = -s^{\perp}(t)$.

The orthonormal basis $\psi_1(t) = \frac{s(t)}{\sqrt{E_s}}$, $\psi_2(t) = \frac{s^{\perp}(t)}{\sqrt{E_s}}$ where $E_s = \int_0^T s_m{}^2(t)\,\mathrm{d}\,t$

$s_1 = \begin{pmatrix} \sqrt{E_s} \\ 0 \end{pmatrix}$, $s_2 = \begin{pmatrix} 0 \\ \sqrt{E_s} \end{pmatrix}$, $s_3 = \begin{pmatrix} -\sqrt{E_s} \\ 0 \end{pmatrix}$, $s_4 = \begin{pmatrix} 0 \\ -\sqrt{E_s} \end{pmatrix}$. The four signals can be geometrically represented using the 4-vector of projection coefficients s_1, s_2, s_3, and s_4 as a set of constellation points.

Signal constellation



Equation:

$$\begin{aligned} d_{21} &= |s_2 - s_1| \\ &= \sqrt{2E_s} \end{aligned}$$

Equation:

$$\begin{aligned} d_{12} &= d_{23} \\ &= d_{34} \\ &= d_{14} \end{aligned}$$

Equation:

$$\begin{aligned} d_{13} &= |s_1 - s_3| \\ &= 2\sqrt{E_s} \end{aligned}$$

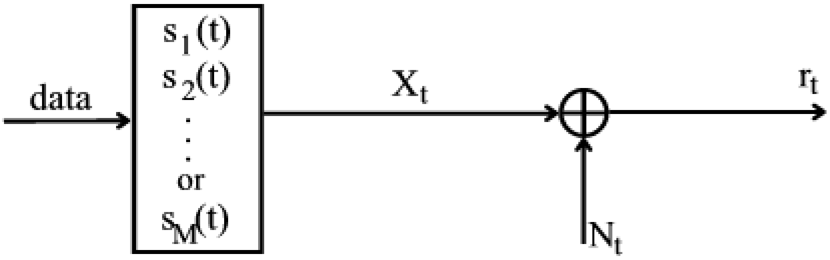
Equation:

$$d_{13} = d_{24}$$

Minimum distance $d_{\text{min}} = \sqrt{2E_s}$

Demodulation and Detection

Consider the problem where signal set, $\{\ _1,\ \ _2, \ldots,\ \ \}$, for $\ \in [0,\ \ ]$ is used to transmit $\log_2$ bits. The **modulated** signal could be $\{\ _1,\ \ _2, \ldots,\ \ \}$ during the interval $0 \le\ \ \le\ \ .$



$$=\quad +\quad =\quad (\ ) +\quad \text{for } 0 \le\ \ \le\ \ \text{for some}\ \ \in \{1, 2, \ldots,\ \ \}.$$

Recall $(\ ) = \ _1 \quad (\ )$ for $\ \in \{1, 2, \ldots,\ \ \}$ the signals are decomposed into a set of orthonormal signals, perfectly.

Noise process can also be decomposed

Equation:

$$=\ {}_1 \quad (\ ) +$$

where $\ = \ _0 \quad (\ )\,\mathrm{d}$ is the projection onto the ${}^{\mathrm{th}}$ basis signal, is the left over noise.

The problem of demodulation and detection is to observe for $0 \le\ \ \le$ and decide which one of the signals were transmitted. Demodulation is covered here. A discussion about detection can be found here.

Demodulation

Demodulation

Convert the continuous time received signal into a vector without loss of information (or performance).

Equation:

$$r_t = s_m(t) + N_t$$

Equation:

$$r_t = \sum_{n=1}^{N} s_{mn}\psi_n(t) + \sum_{n=1}^{N} \eta_n\psi_n(t) + \widetilde{N_t}$$

Equation:

$$r_t = \sum_{n=1}^{N} (s_{mn} + \eta_n)\psi_n(t) + \widetilde{N_t}$$

Equation:

$$r_t = \sum_{n=1}^{N} r_n\psi_n(t) + \widetilde{N_t}$$

The noise projection coefficients η_n's are zero mean, Gaussian random variables and are mutually independent if N_t is a white Gaussian process.

Equation:

$$\begin{array}{rcl} \mu_\eta(n) & = & E[\eta_n] \\ & = & E\left[\int_0^T N_t\psi_n(t)\,\mathrm{d}\,t\right] \end{array}$$

Equation:

$$\begin{array}{rcl} \mu_\eta(n) & = & \int_0^T E[N_t]\psi_n(t) \text{ d } t \\ & = & 0 \end{array}$$

Equation:

$$\begin{array}{rcl} E[\eta_k\overline{\eta_n}] & = & E\left[\int_0^T N_t\psi_k(t) \text{ d } t \int_0^T \overline{N_{t'}}\overline{\psi_k(t')} \text{ d } t'\right] \\ & = & \int_0^T \int_0^T \overline{N_t N_{t'}}\psi_k(t)\psi_n(t') \text{ d } t \text{ d } t' \end{array}$$

Equation:

$$E[\eta_k\overline{\eta_n}] = \int_0^T \int_0^T R_N\big(t-t'\big)\psi_k(t)\overline{\psi_n} \text{ d } t \text{ d } t'$$

Equation:

$$E[\eta_k\overline{\eta_n}] = \frac{N_0}{2}\int_0^T \int_0^T \delta\big(t-t'\big)\psi_k(t)\overline{\psi_n(t')} \text{ d } t \text{ d } t'$$

Equation:

$$\begin{array}{rcl} E[\eta_k\overline{\eta_n}] & = & \frac{N_0}{2}\int_0^T \psi_k(t)\overline{\psi_n(t)} \text{ d } t \\ & = & \frac{N_0}{2}\delta_{kn} \\ & = & \begin{cases} \frac{N_0}{2} \text{ if } k = n \\ 0 \text{ if } k \neq n \end{cases} \end{array}$$

η_k 's are uncorrelated and since they are Gaussian they are also independent. Therefore, $\eta_k \simeq \text{Gaussian}\left(0, \frac{N_0}{2}\right)$ and $R_\eta(k,n) = \frac{N_0}{2}\delta_{kn}$

The r_n's, the projection of the received signal r_t onto the orthonormal bases $\psi_n(t)$'s, are independent from the residual noise process $\widetilde{N_t}$.

The residual noise $\widetilde{N_t}$ is irrelevant to the decision process on r_t.

Recall $r_n = s_{mn} + \eta_n$, given $s_m(t)$ was transmitted. Therefore,

Equation:

$$\begin{array}{rcl} \mu_r(n) & = & E[s_{mn} + \eta_n] \\ & = & s_{mn} \end{array}$$

Equation:

$$\begin{array}{rcl} \mathrm{Var}\,(r_n) & = & \mathrm{Var}\,(\eta_n) \\ & = & \frac{N_0}{2} \end{array}$$

The correlation between r_n and $\widetilde{N_t}$

Equation:

$$E\left[\widetilde{N_t}\overline{r_n}\right] = E\left[\left(N_t - \sum_{k=1}^{N} \eta_k \psi_k(t)\right)\overline{s_{mn} + \eta_n}\right]$$

Equation:

$$E\left[\widetilde{N_t}\overline{r_n}\right] = E\left[N_t - \sum_{k=1}^{N} \eta_k \psi_k(t)\right] s_{mn} + E[\eta_k \overline{\eta_n}] - \sum_{k=1}^{N} E[\eta_k \overline{\eta_n}]\psi_k(t)$$

Equation:

$$E\left[\widetilde{N_t}\overline{r_n}\right] = E\left[N_t \int_0^T \overline{N_{t'}}\overline{\psi_n(t')}\,\mathrm{d}\,t'\right] - \sum_{k=1}^{N} \frac{N_0}{2}\delta_{kn}\psi_k(t)$$

Equation:

$$E\left[\widetilde{N_t}\overline{r_n}\right] = \int_0^T \frac{N_0}{2}\delta(t - t')\psi_n(t')\,\mathrm{d}\,t' - \frac{N_0}{2}\psi_n(t)$$

Equation:

$$
\begin{aligned}
E\left[\widetilde{N_t}\overline{r_n}\right] &= \tfrac{N_0}{2}\psi_n(t) - \tfrac{N_0}{2}\psi_n(t) \\
&= 0
\end{aligned}
$$

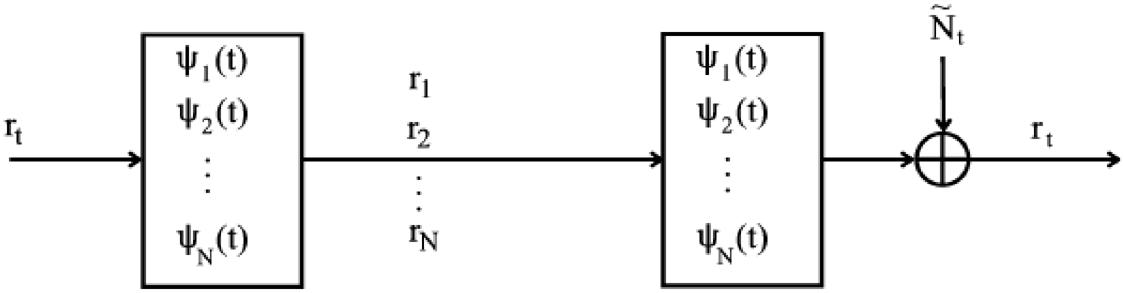
Since both $\widetilde{N_t}$ and r_n are Gaussian then $\widetilde{N_t}$ and r_n are also independent.

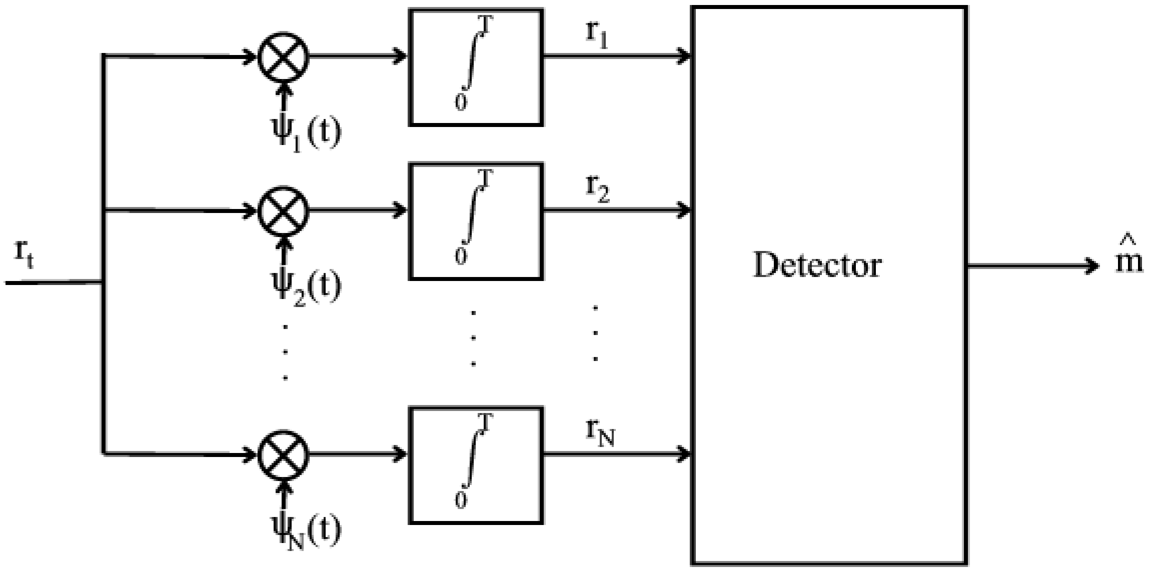
The conjecture is to ignore $\widetilde{N_t}$ and extract information from $\begin{pmatrix} r_1 \\ r_2 \\ \dots \\ r_N \end{pmatrix}$.

Knowing the vector $\boldsymbol{r}$ we can reconstruct the relevant part of random process r_t for $0 \leq t \leq T$

Equation:

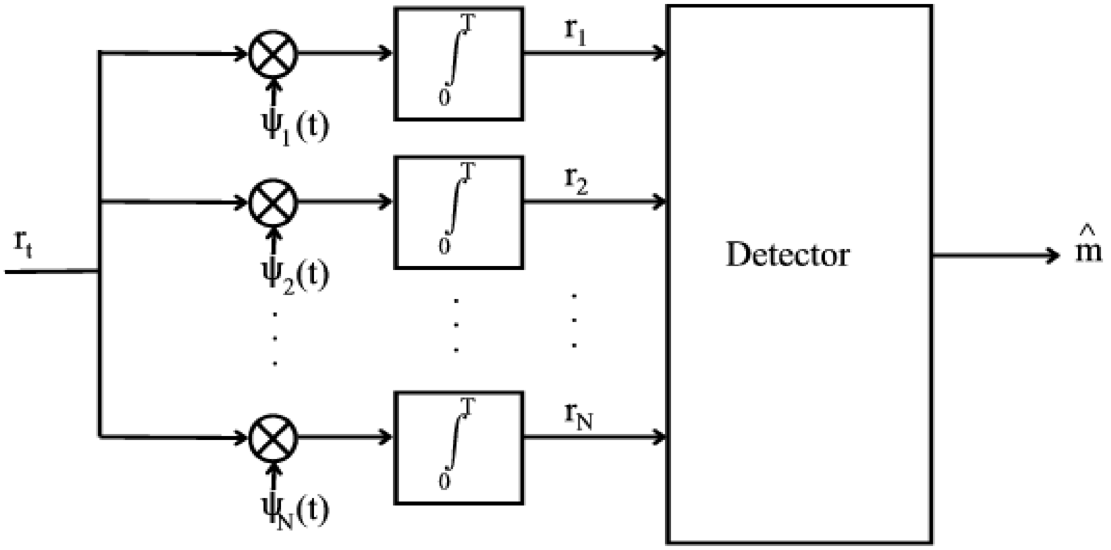
$$
\begin{aligned}
r_t &= s_m(t) + N_t \\
&= \textstyle\sum_{n=1}^{N} r_n\psi_n(t) + \widetilde{N_t}
\end{aligned}
$$





Once the received signal has been converted to a vector, the correct transmitted signal must be detected based upon observations of the input vector. Detection is covered elsewhere.

Detection by Correlation
Demodulation and Detection



Detection

Decide which $s_m(t)$ from the set of $\{s_1(t), \ldots, s_m(t)\}$ signals was transmitted based on observing $\boldsymbol{r} = \begin{matrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_N \end{matrix}$, the vector composed of demodulated received signal, that is, the vector of projection of the received signal onto the N bases.

Equation:

$$\hat{m} = \arg \max_{1 \le m \le M} \Pr[s_m(t) \text{ was transmitted } \mid \mathbf{r} \text{ was observed}]$$

Note that

Equation:

$$\Pr[s_m \mid \boldsymbol{r}] \triangleq \Pr[\boldsymbol{s}_m(t) \text{was transmitted} \mid \mathbf{r} \text{ was observed}] = \frac{f_{r|\boldsymbol{s}_m} \Pr[\boldsymbol{s}_m]}{f_{\boldsymbol{r}}}$$

If $\Pr[\boldsymbol{s}_m \text{ was transmitted}] = \frac{1}{M}$, that is information symbols are equally likely to be transmitted, then

Equation:

$$\arg\max_{1 \le m \le M} \Pr[\boldsymbol{s}_m \mid \boldsymbol{r}] = \arg\max_{1 \le m \le M} f_{\boldsymbol{r}|\boldsymbol{s}_m}$$

Since $r(t) = s_m(t) + N_t$ for $0 \le t \le T$ and for some $m = \{1, 2, \ldots, M\}$

then $\boldsymbol{r} = \boldsymbol{s}_m + \boldsymbol{\eta}$ where $\boldsymbol{\eta} = \begin{matrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \vdots \\ \boldsymbol{\eta}_N \end{matrix}$ and $\boldsymbol{\eta}_n$'s are Gaussian and independent.

Equation:

$$\forall \boldsymbol{r}_n, \boldsymbol{r}_n \in \mathbb{R} : \quad f_{\boldsymbol{r}|\boldsymbol{s}_m} = \frac{1}{\left(2\pi \frac{N_0}{2}\right)^{\frac{N}{2}}} e^{\frac{-\sum_{n=1}^{N} (\boldsymbol{r}_n - s_{m,n})^2}{2\frac{N_0}{2}}}$$

Equation:

$$\begin{aligned} \widehat{m} &= \arg\max_{1 \le m \le M} f_{\boldsymbol{r}|\boldsymbol{s}_m} \\ &= \arg\max_{1 \le m \le M} \ln\left(f_{\boldsymbol{r}|\boldsymbol{s}_m}\right) \\ &= \arg\max_{1 \le m \le M} \left(-\left(\frac{N}{2} \ln(\pi N_0)\right)\right) - \frac{1}{N_0} \sum_{n=1}^{N} (\boldsymbol{r}_n - s_{m,n})^2 \\ &= \arg\min_{1 \le m \le M} \sum_{n=1}^{N} (\boldsymbol{r}_n - s_{m,n})^2 \end{aligned}$$

where $D(\boldsymbol{r}, \boldsymbol{s}_m)$ is the l_2 distance between vectors $\boldsymbol{r}$ and $\boldsymbol{s}_m$ defined as $D(\boldsymbol{r}, \boldsymbol{s}_m) \triangleq \sum_{n=1}^{N} (\boldsymbol{r}_n - s_{m,n})^2$

Equation:

$$\begin{aligned}\widehat{m} &= \arg\min_{1\leq m\leq M} D(\boldsymbol{r}, \boldsymbol{s}_m) \\ &= \arg\min_{1\leq m\leq M} (\| \boldsymbol{r} \|)^2 - 2\left\langle(\boldsymbol{r}, \boldsymbol{s}_m)\right\rangle + (\| \boldsymbol{s}_m \|)^2\end{aligned}$$

where $\| \boldsymbol{r} \|$ is the l_2 norm of vector $\boldsymbol{r}$ defined as $\| \boldsymbol{r} \| \triangleq \sqrt{\sum_{n=1}^{N} (\boldsymbol{r}_n)^2}$

Equation:

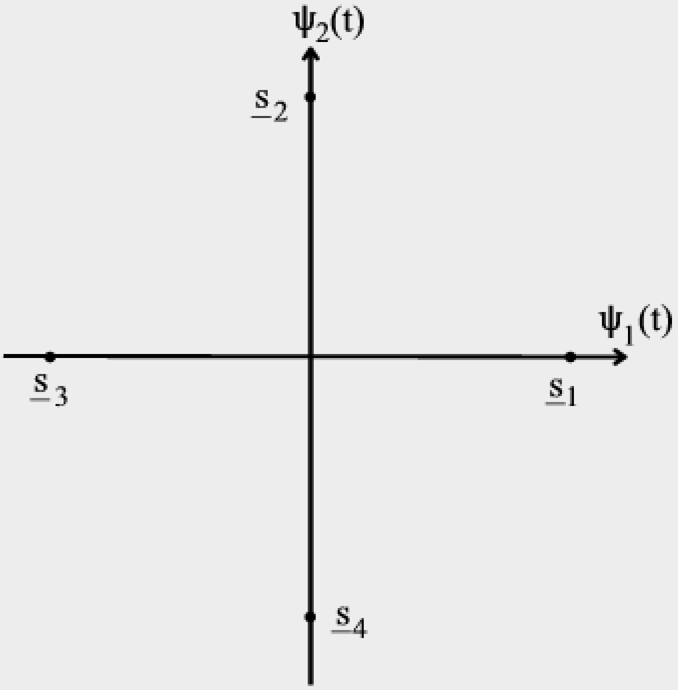
$$\widehat{m} = \arg\max_{1\leq m\leq M} 2\left\langle(\boldsymbol{r}, \boldsymbol{s}_m)\right\rangle - (\| \boldsymbol{s}_m \|)^2$$

This type of receiver system is known as a **correlation** (or correlator-type) receiver. Examples of the use of such a system are found here. Another type of receiver involves linear, time-invariant filters and is known as a matched filter receiver. An analysis of the performance of a correlator-type receiver using antipodal and orthogonal binary signals can be found in Performance Analysis.

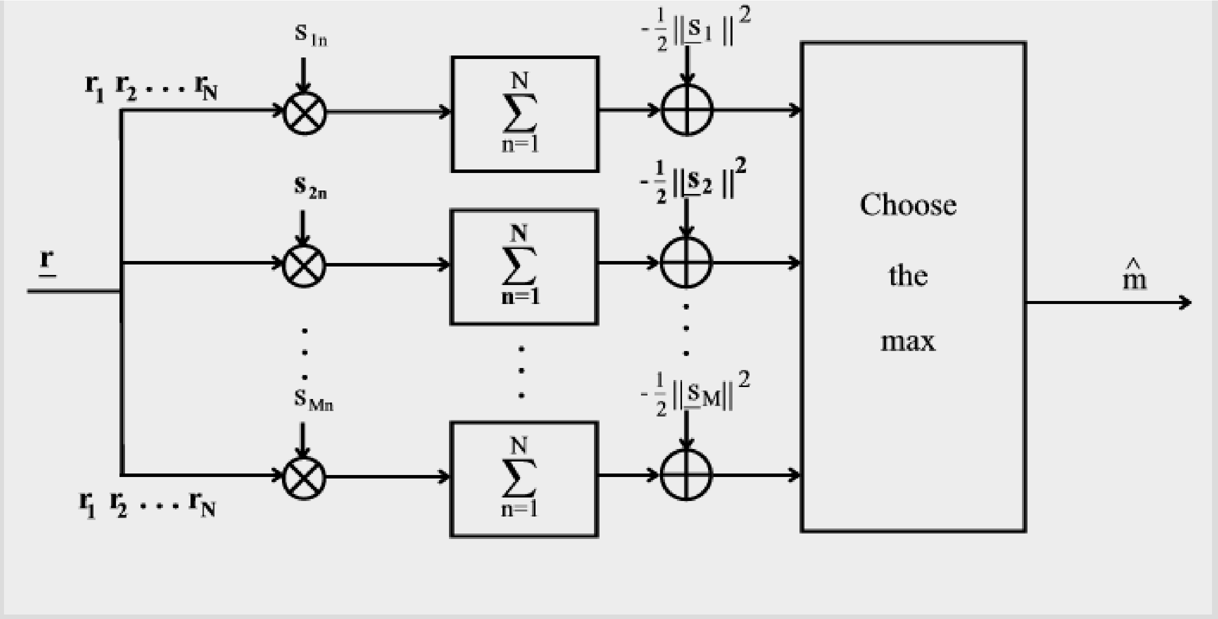
Examples of Correlation Detection

The implementation and theory of correlator-type receivers can be found in Detection.

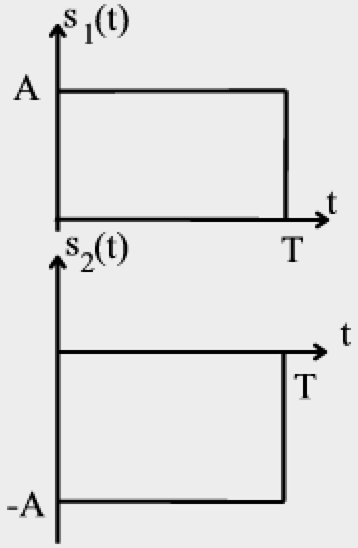
Example:



$\widehat{m} = 2$ since $D(\boldsymbol{r}, s_1) > D(\boldsymbol{r}, s_2)$ or $(\| s_1 \|)^2 = (\| s_2 \|)^2$ and $\langle \boldsymbol{r}, s_2 \rangle > \langle \boldsymbol{r}, s_1 \rangle$.



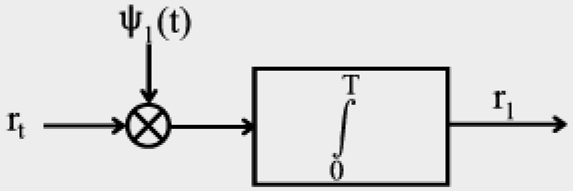
Example:
Data symbols "0" or "1" with equal probability. Modulator $s_1(t) = s(t)$ for $0 \leq t \leq T$ and $s_2(t) = -s(t)$ for $0 \leq t \leq T$.



$\psi_1(t) = \frac{s(t)}{\sqrt{A^2 T}}$, $s_{11} = A\sqrt{T}$, and $s_{21} = -\left(A\sqrt{T}\right)$

Equation:

$$\forall m, m = \{1, 2\} : (r_t = s_m(t) + N_t)$$



Equation:

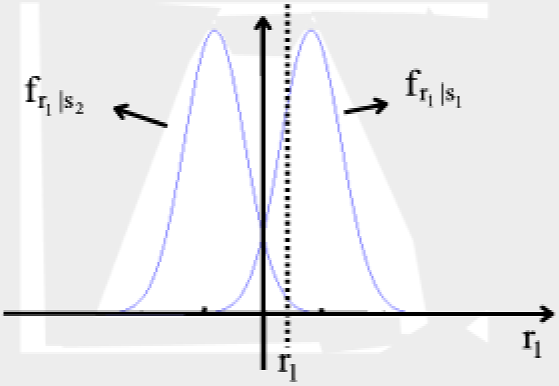
$$r_1 = A\sqrt{T} + \eta_1$$

or

Equation:

$$r_1 = -\left(A\sqrt{T}\right) + \eta_1$$

η_1 is Gaussian with zero mean and variance $\frac{N_0}{2}$.



$\widehat{m}$ =argmax $\left\{A\sqrt{T}r_1, -\left(A\sqrt{T}r_1\right)\right\}$, since $A\sqrt{T} > 0$ and $\Pr[s_1] = \Pr[s_1]$ then the MAP decision rule decides.

$s_1(t)$ was transmitted if $r_1 \geq 0$

$s_2(t)$ was transmitted if $r_1 < 0$

An alternate demodulator:

Equation:

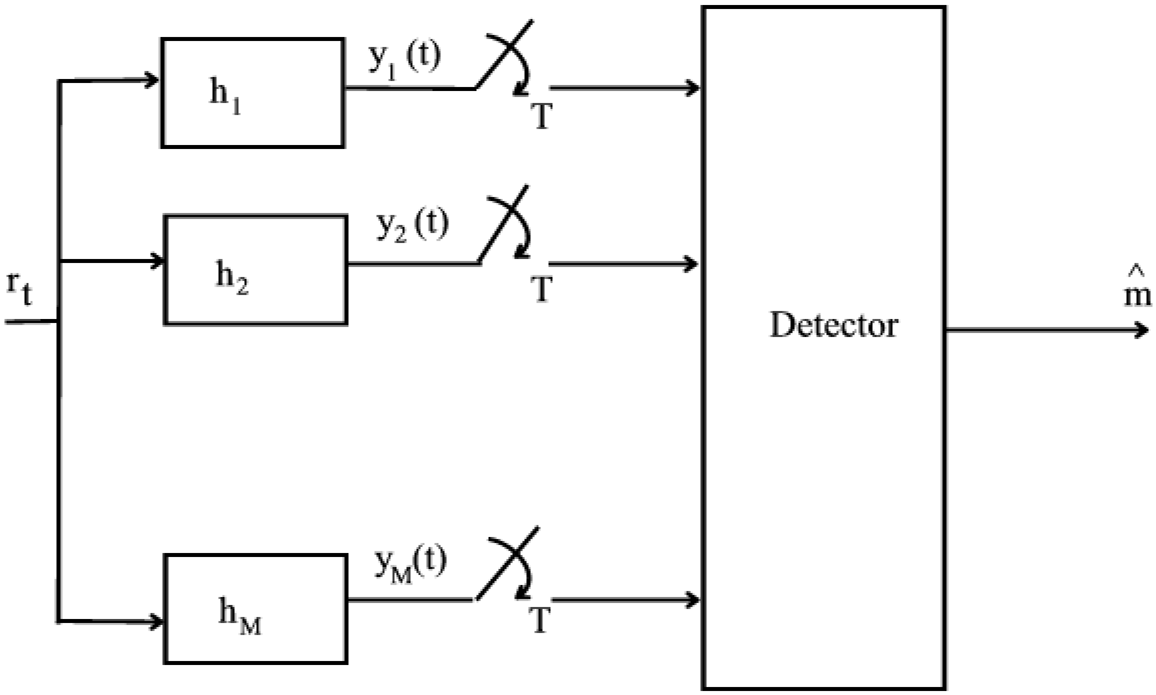
$$(r_t = s_m(t) + N_t) \Rightarrow (\boldsymbol{r} = s_m + \eta)$$

Matched Filters

Signal to Noise Ratio (SNR) at the output of the demodulator is a measure of the quality of the demodulator.
Equation:

$$\mathrm{SNR} = \frac{\text{signal energy}}{\text{noise energy}}$$

In the correlator described earlier, $E_s = (|s_m|)^2$ and ${\sigma_{\eta_n}}^2 = \frac{N_0}{2}$. Is it possible to design a demodulator based on linear time-invariant filters with maximum signal-to-noise ratio?



If $s_m(t)$ is the transmitted signal, then the output of the k^{th} filter is given as
Equation:

$$\begin{aligned} y_k(t) &= \int_{-\infty}^{\infty} r_\tau h_k(t-\tau)\,\mathrm{d}\,\tau \\ &= \int_{-\infty}^{\infty} (s_m(\tau)+N_\tau)h_k(t-\tau)\,\mathrm{d}\,\tau \\ &= \int_{-\infty}^{\infty} s_m(\tau)h_k(t-\tau)\,\mathrm{d}\,\tau + \int_{-\infty}^{\infty} N_\tau h_k(t-\tau)\,\mathrm{d}\,\tau \end{aligned}$$

Sampling the output at time T yields
Equation:

$$y_k(T) = \int_{-\infty}^{\infty} s_m(\tau)h_k(T-\tau)\,\mathrm{d}\,\tau + \int_{-\infty}^{\infty} N_\tau h_k(T-\tau)\,\mathrm{d}\,\tau$$

The noise contribution:
Equation:

$$\nu_k = \int_{-\infty}^{\infty} N_\tau h_k(T-\tau)\,\mathrm{d}\,\tau$$

The expected value of the noise component is
Equation:

$$\begin{aligned} E[\nu_k] &= E\left[\int_{-\infty}^{\infty} N_\tau h_k(T-\tau)\,\mathrm{d}\,\tau\right] \\ &= 0 \end{aligned}$$

The variance of the noise component is the second moment since the mean is zero and is given as
Equation:

$$\begin{aligned} \sigma(\nu_k)^2 &= E\left[{\nu_k}^2\right] \\ &= E\left[\int_{-\infty}^{\infty} N_\tau h_k(T-\tau)\,\mathrm{d}\,\tau \int_{-\infty}^{\infty} N_{\tau'} h_k(T-\tau')\,\mathrm{d}\,\tau'\right] \end{aligned}$$

Equation:

$$\begin{aligned} E\left[{\nu_k}^2\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tfrac{N_0}{2}\delta\left(\tau-\tau\text{'}\right)h_k(T-\tau)h_k(T-\tau\text{'})\ \mathrm{d}\,\tau\ \mathrm{d}\,\tau\text{'} \\ &= \tfrac{N_0}{2}\int_{-\infty}^{\infty}\left(\left|h_k(T-\tau)\right|\right)^2\ \mathrm{d}\,\tau \end{aligned}$$

Signal Energy can be written as

Equation:

$$\left(\int_{-\infty}^{\infty}s_m(\tau)h_k(T-\tau)\ \mathrm{d}\,\tau\right)^2$$

and the signal-to-noise ratio (SNR) as

Equation:

$$\mathrm{SNR}=\frac{\left(\int_{-\infty}^{\infty}s_m(\tau)h_k(T-\tau)\ \mathrm{d}\,\tau\right)^2}{\frac{N_0}{2}\int_{-\infty}^{\infty}\left(\left|h_k(T-\tau)\right|\right)^2\ \mathrm{d}\,\tau}$$

The signal-to-noise ratio, can be maximized considering the well-known Cauchy-Schwarz Inequality

Equation:

$$\left(\int_{-\infty}^{\infty}g_1(x)g_2(x)\ \mathrm{d}\,x\right)^2\le\int_{-\infty}^{\infty}\left(\left|g_1(x)\right|\right)^2\ \mathrm{d}\,x\int_{-\infty}^{\infty}\left(\left|g_2(x)\right|\right)^2\ \mathrm{d}\,x$$

with equality when $g_1(x)=\alpha g_2(x)$. Applying the inequality directly yields an upper bound on SNR

Equation:

$$\frac{\left(\int_{-\infty}^{\infty}s_m(\tau)h_k(T-\tau)\ \mathrm{d}\,\tau\right)^2}{\frac{N_0}{2}\int_{-\infty}^{\infty}\left(\left|h_k(T-\tau)\right|\right)^2\ \mathrm{d}\,\tau}\le\frac{2}{N_0}\int_{-\infty}^{\infty}\left(\left|s_m(\tau)\right|\right)^2\ \mathrm{d}\,\tau$$

with equality $\forall \tau : \left(h_k^{\text{opt}}(T-\tau) = \alpha s_m(\tau)\right)$. Therefore, the filter to examine signal m should be

Equation:

Matched Filter

$$\forall \tau : \left(h_m^{\text{opt}}(\tau) = s_m(T-\tau)\right)$$

The constant factor is not relevant when one considers the signal to noise ratio. The maximum SNR is unchanged when both the numerator and denominator are scaled.

Equation:

$$\frac{2}{N_0}\int_{-\infty}^{\infty} \left(|s_m(\tau)|\right)^2 \,\mathrm{d}\,\tau = \frac{2E_s}{N_0}$$

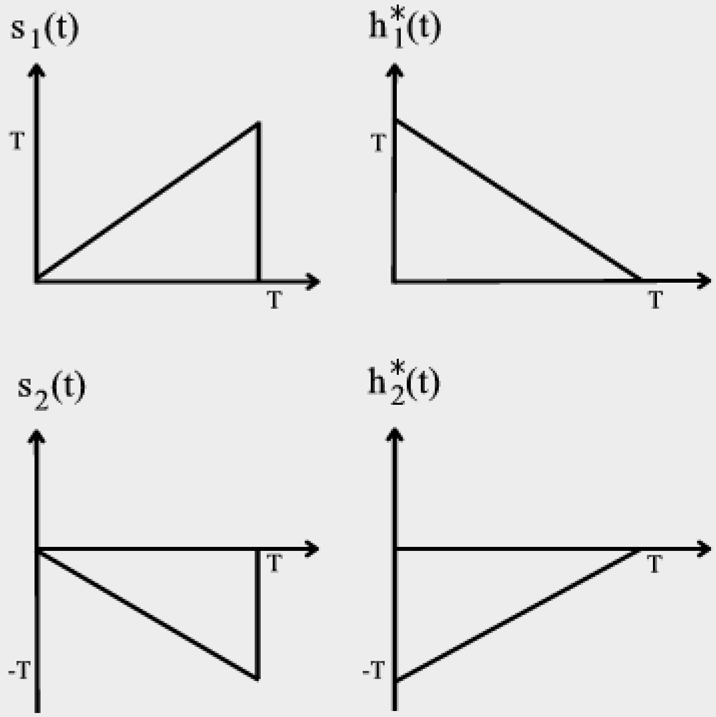
Examples involving matched filter receivers can be found here. An analysis in the frequency domain is contained in Matched Filters in the Frequency Domain.

Another type of receiver system is the correlation receiver. A performance analysis of both matched filters and correlator-type receivers can be found in Performance Analysis.

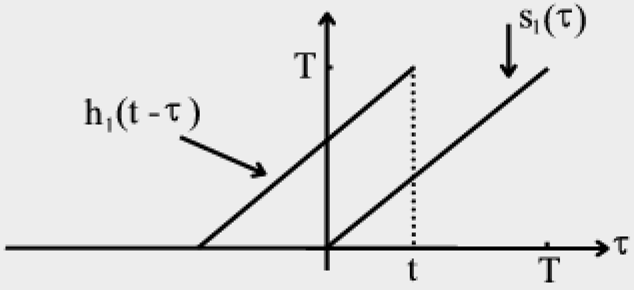
Examples with Matched Filters

The theory and rationale behind matched filter receivers can be found in Matched Filters.

Example:



$s_1(t) = t$ for $0 \leq t \leq T$
$s_2(t) = -t$ for $0 \leq t \leq T$
$h_1(t) = T - t$ for $0 \leq t \leq T$
$h_2(t) = -T + t$ for $0 \leq t \leq T$



Equation:

$$\forall t, 0 \leq t \leq 2T : \left(\widetilde{s_1}(t) = \int_{-\infty}^{\infty} s_1(\tau) h_1(t-\tau) \, \mathrm{d}\, \tau\right)$$

Equation:

$$\begin{aligned} \widetilde{s_1}(t) &= \int_0^t \tau \,(T - t + \tau) \, \mathrm{d}\, \tau \\ &= \tfrac{1}{2}\,(T-t)\tau^2\big|_0^t + \tfrac{1}{3}\tau^3\big|_0^t \\ &= \tfrac{t^2}{2}\left(T - \tfrac{t}{3}\right) \end{aligned}$$

Equation:

$$\widetilde{s_1}(T) = \frac{T^3}{3}$$

Compared to the correlator-type demodulation

Equation:

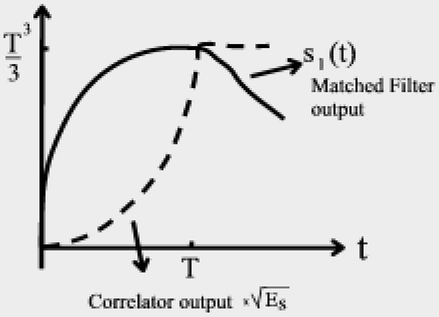
$$\psi_1(t) = \frac{s_1(t)}{\sqrt{E_s}}$$

Equation:

$$s_{11} = \int_0^T s_1(\tau)\psi_1(\tau) \, \mathrm{d}\, \tau$$

Equation:

$$\begin{aligned} \int_0^t s_1(\tau)\psi_1(\tau) \, \mathrm{d}\, \tau &= \tfrac{1}{\sqrt{E_s}} \int_0^t \tau\tau \, \mathrm{d}\, \tau \\ &= \tfrac{1}{\sqrt{E_s}} \tfrac{1}{3} t^3 \end{aligned}$$



Example:

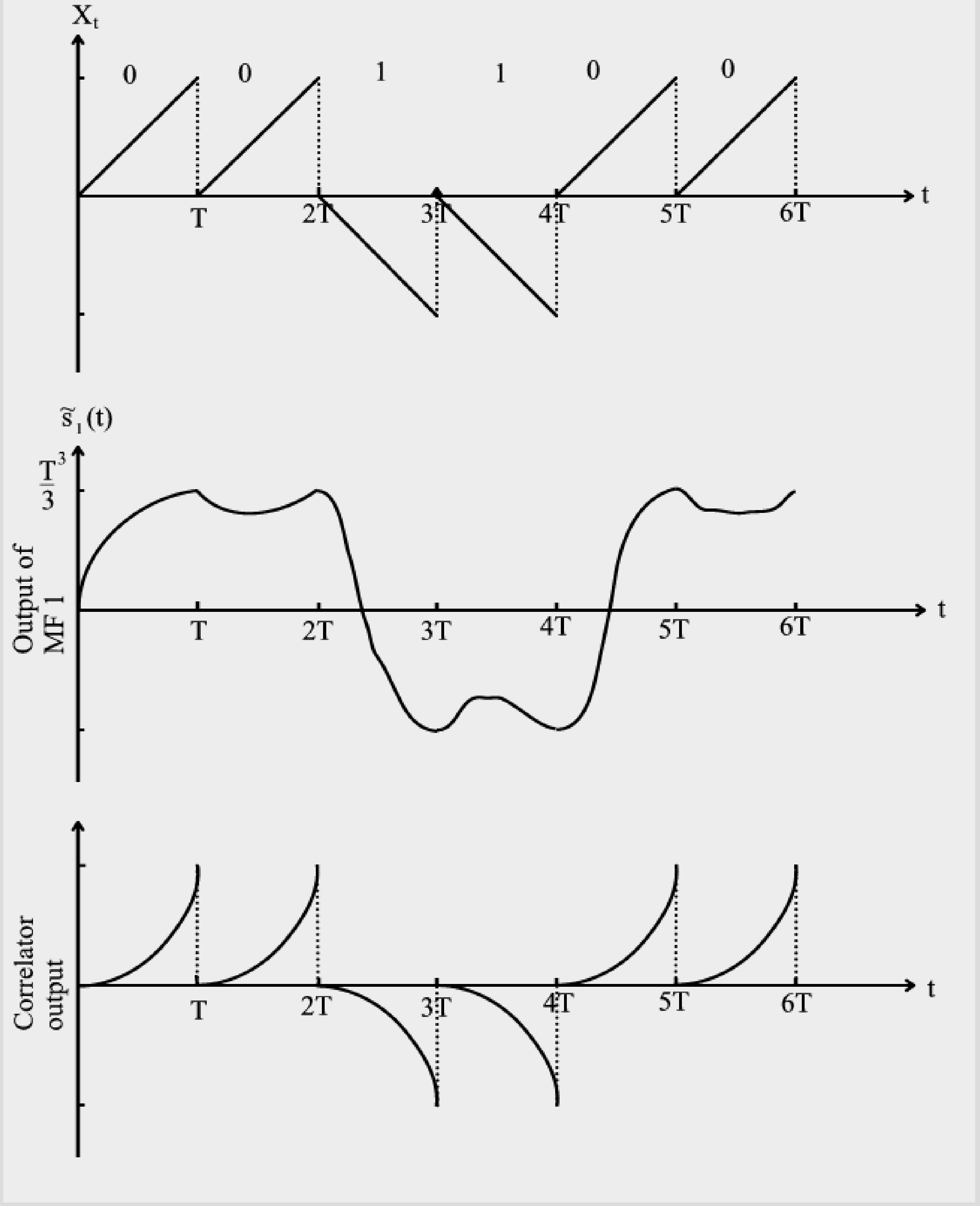
Assume binary data is transmitted at the rate of $\frac{1}{T}$ Hertz.

$0 \Rightarrow (b = 1) \Rightarrow (s_1(t) = s(t))$ for $0 \leq t \leq T$

$1 \Rightarrow (b = -1) \Rightarrow (s_2(t) = -s(t)) \quad \text{for} \quad 0 \leq t \leq T$

Equation:

$$X_t = \sum_{i \quad -P}^{P} b_i s(t - iT)$$

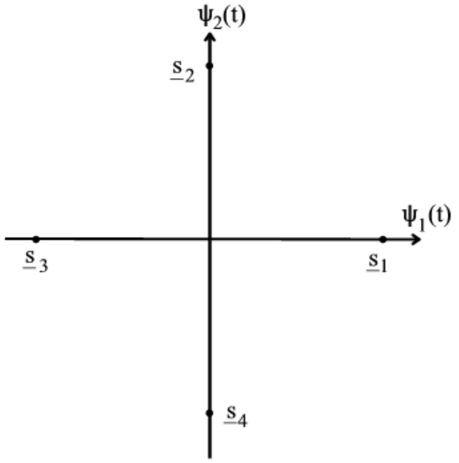

X_t
0
0
1
1
0
0
T
2T
3T
4T
5T
6T
t
$\tilde{s}_1(t)$
$\frac{T^3}{3}$
Output of MF 1
T
2T
3T
4T
5T
6T
t
Correlator output
T
2T
3T
4T
5T
6T
t

Performance Analysis of Binary Orthogonal Signals with Correlation

Orthogonal signals with equally likely bits, $r_t = s_m(t) + N_t$ for $0 \leq t \leq T$, $m = 1$, $m = 2$, and $\langle s_1, s_2 \rangle = 0$.

Correlation (correlator-type) receiver

$r_t \Rightarrow \left(\boldsymbol{r} = (r_1 r_2)^T = \boldsymbol{s}_m + \boldsymbol{\eta}\right)$ (see [link])



Decide $s_1(t)$ was transmitted if $r_1 \geq r_2$.

Equation:

$$\begin{aligned} P_e &= \Pr\left[\widehat{m} \neq m\right] \\ &= \Pr\left[\hat{b} \neq b\right] \end{aligned}$$

Equation:

$$\begin{aligned} P_e &= 1/2 \Pr[\boldsymbol{r} \in R_2 \mid s_1(t) \text{ transmitted}] + 1/2 \Pr[\boldsymbol{r} \in R_1 \mid s_2(t) \text{ transmitted}] \\ &= 1/2 \int_{R_2} \int \mathrm{f}_{\boldsymbol{r}, s_1(t)}\,(\boldsymbol{r})\;\mathrm{d}\,r_1\;\mathrm{d}\,r_2 + 1/2 \int_{R_1} \int \mathrm{f}_{\boldsymbol{r}, s_2(t)}\,(\boldsymbol{r})\;\mathrm{d}\,r_1\;\mathrm{d}\,r_2 \\ &= 1/2 \int_{R_2} \int \frac{1}{\sqrt{2\pi \frac{N_0}{2}}} e^{\frac{-\left(\left|r_1 - \sqrt{E_s}\right|\right)^2}{N_0}} \frac{1}{\sqrt{\pi N_0}} e^{\frac{-\left(\left|r_2\right|\right)^2}{N_0}}\;\mathrm{d}\,r_1\;\mathrm{d}\,r_2 + 1/2 \int_{R_1} \int \frac{1}{\sqrt{2\pi \frac{N_0}{2}}} e^{\frac{-\left(\left|r_1\right|\right)^2}{N_0}} \frac{1}{\sqrt{\pi N_0}} e^{\frac{-\left(\left|r_2 - \sqrt{E_s}\right|\right)^2}{N_0}}\;\mathrm{d}\,r_1\;\mathrm{d} \end{aligned}$$

Alternatively, if $s_1(t)$ is transmitted we decide on the wrong signal if $r_2 > r_1$ or $\eta_2 > \eta_1 + \sqrt{E_s}$ or when $\eta_2 - \eta_1 > \sqrt{E_s}$.

Equation:

$$\begin{aligned} P_e &= 1/2 \int_{\sqrt{E_s}}^{\infty} \frac{1}{\sqrt{2\pi N_0}} e^{\frac{-\eta'^2}{2N_0}}\;\mathrm{d}\,\eta' + 1/2 \Pr[r_1 \geq r_2 \mid s_2(t) \text{ transmitted}] \\ &= Q\left(\sqrt{\frac{E_s}{N_0}}\right) \end{aligned}$$

Note that the distance between s_1 and s_2 is $d_{12} = \sqrt{2E_s}$. The average bit error probability $P_e = Q\left(\frac{d_{12}}{\sqrt{2N_0}}\right)$ as we had for the antipodal case. Note also that the bit-error probability is the same as for the matched filter receiver.

Performance Analysis of Orthogonal Binary Signals with Matched Filters

Equation:

$$r_t \Rightarrow \left(\boldsymbol{Y} = \begin{pmatrix} Y_1(T) \\ Y_2(T) \end{pmatrix}\right)$$

If $s_1(t)$ is transmitted

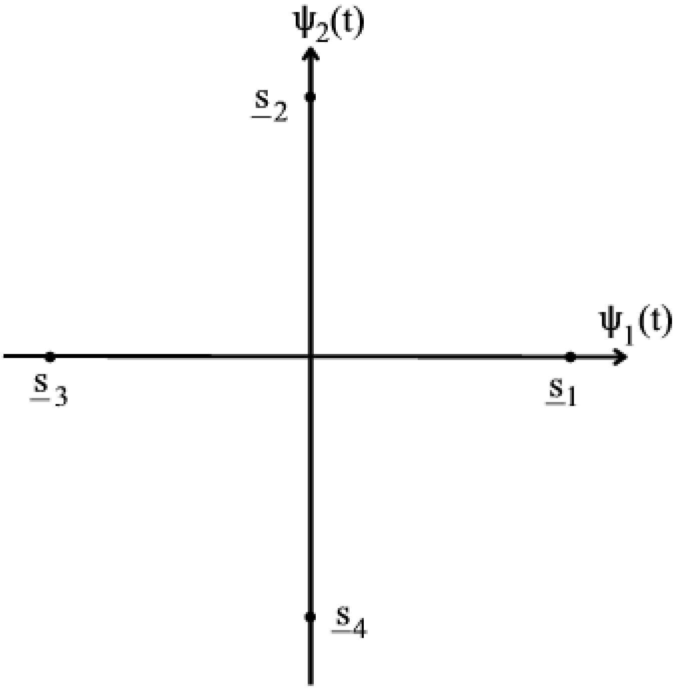
Equation:

$$\begin{array}{rcl} Y_1(T) & = & \int_{-\infty}^{\infty} s_1(\tau) h_1^{\mathrm{opt}}(T-\tau)\,\mathrm{d}\,\tau + \nu_1(T) \\ & = & \int_{-\infty}^{\infty} s_1(\tau) \overset{*}{s_1}(\tau)\,\mathrm{d}\,\tau + \nu_1(T) \\ & = & E_s + \nu_1(T) \end{array}$$

Equation:

$$\begin{array}{rcl} Y_2(T) & = & \int_{-\infty}^{\infty} s_1(\tau) \overset{*}{s_2}(\tau)\,\mathrm{d}\,\tau + \nu_2(T) \\ & = & \nu_2(T) \end{array}$$

If $s_2(t)$ is transmitted, $Y_1(T) = \nu_1(T)$ and $Y_2(T) = E_s + \nu_2(T)$.



Equation:

H0

$$\boldsymbol{Y} = \begin{pmatrix} E_s \\ 0 \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$$

Equation:

H1

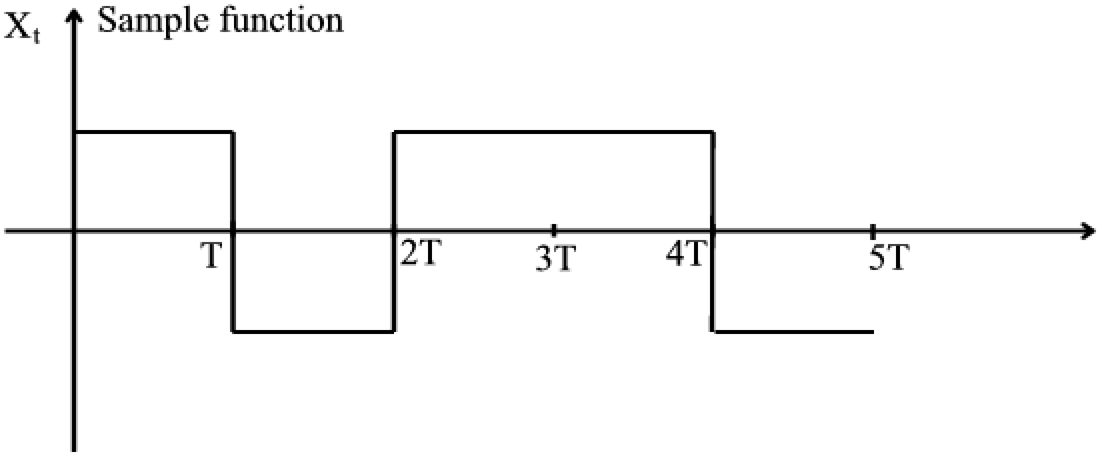
$$\boldsymbol{Y} = \begin{pmatrix} 0 \\ E_s \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$$

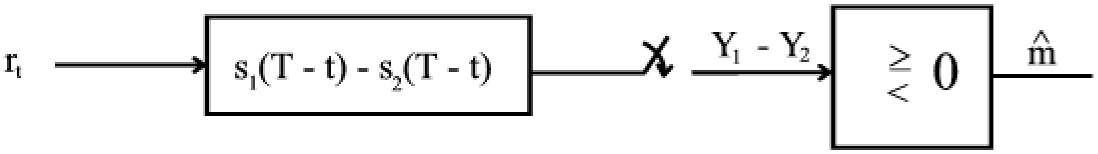
where ν_1 and ν_2 are independent are Gaussian with zero mean and variance $\frac{N_0}{2}E_s$. The analysis is identical to the correlator example.

Equation:

$$P_e = Q\left(\sqrt{\frac{E_s}{N_0}}\right)$$

Note that the maximum likelihood detector decides based on comparing Y_1 and Y_2. If $Y_1 \geq Y_2$ then s_1 was sent; otherwise s_2 was transmitted. For a similar analysis for binary antipodal signals, refer here. See [link] or [link].





Carrier Phase Modulation

Phase Shift Keying (PSK)

Information is impressed on the phase of the carrier. As data changes from symbol period to symbol period, the phase shifts.

Equation:

$$\forall m, m \in \{1, 2, \ldots, M\} : \left(s_m(t) = AP_T(t)\cos\left(2\pi f_c t + \frac{2\pi\,(m-1)}{M}\right)\right)$$

Example:
Binary $s_1(t)$ or $s_2(t)$

Representing the Signals

An orthonormal basis to represent the signals is

Equation:

$$\psi_1(t) = \frac{1}{\sqrt{E_s}} AP_T(t)\cos(2\pi f_c t)$$

Equation:

$$\psi_2(t) = \frac{-1}{\sqrt{E_s}} AP_T(t)\sin(2\pi f_c t)$$

The signal

Equation:

$$S_m(t) = AP_T(t)\cos\left(2\pi f_c t + \frac{2\pi\,(m-1)}{M}\right)$$

Equation:

$$S_m(t) = A\cos\left(\frac{2\pi\,(m-1)}{M}\right)P_T(t)\cos(2\pi f_c t) - A\sin\left(\frac{2\pi\,(m-1)}{M}\right)P_T(t)\sin(2\pi f_c t)$$

The signal energy

Equation:

$$\begin{aligned} E_s &= \int_{-\infty}^{\infty} A^2 {P_T}^2(t)\cos^2\left(2\pi f_c t + \tfrac{2\pi(m-1)}{M}\right)\,\mathrm{d}\,t \\ &= \int_0^T A^2\left(\tfrac{1}{2} + \tfrac{1}{2}\cos\left(4\pi f_c t + \tfrac{4\pi(m-1)}{M}\right)\right)\,\mathrm{d}\,t \end{aligned}$$

Equation:

$$E_s = \frac{A^2T}{2} + \frac{1}{2}A^2\int_0^T \cos\left(4\pi f_c t + \frac{4\pi\,(m-1)}{M}\right)\,\mathrm{d}\,t \simeq \frac{A^2T}{2}$$

(Note that in the above equation, the integral in the last step before the aproximation is very small.) Therefore,
Equation:

$$\psi_1(t) = \sqrt{\frac{2}{T}}P_T(t)\cos(2\pi f_c t)$$

Equation:

$$\psi_2(t) = \left(-\sqrt{\frac{2}{T}}\right)P_T(t)\sin(2\pi f_c t)$$

In general,
Equation:

$$\forall m, m \in \{1, 2, \ldots, M\} : \left(s_m(t) = AP_T(t)\cos\left(2\pi f_c t + \frac{2\pi\,(m-1)}{M}\right)\right)$$

and $\psi_1(t)$
Equation:

$$\psi_1(t) = \sqrt{\frac{2}{T}}P_T(t)\cos(2\pi f_c t)$$

Equation:

$$\psi_2(t) = \sqrt{\frac{2}{T}}P_T(t)\sin(2\pi f_c t)$$

Equation:

$$s_m = \begin{pmatrix} \sqrt{E_s}\cos\left(\frac{2\pi(m-1)}{M}\right) \\ \sqrt{E_s}\sin\left(\frac{2\pi(m-1)}{M}\right) \end{pmatrix}$$

Demodulation and Detection

Equation:

$$r_t = s_m(t) + N_t, \text{ for some} m \in \{1, 2, \ldots, M\}$$

We must note that due to phase offset of the oscillator at the transmitter, **phase jitter** or **phase changes** occur because of propagation delay.
Equation:

$$r_t = AP_T(t)\cos\left(2\pi f_c t + \frac{2\pi\,(m-1)}{M} + \varphi\right) + N_t$$

For binary PSK, the modulation is antipodal, and the optimum receiver in AWGN has average bit-error probability

Equation:

$$\begin{array}{rcl} P_e & = & Q\left(\sqrt{\frac{2(E_s)}{N_0}}\right) \\ & = & Q\left(A\sqrt{\frac{T}{N_0}}\right) \end{array}$$

The receiver where

Equation:

$$r_t = \pm(AP_T(t)\cos(2\pi f_c t + \varphi)) + N_t$$

The statistics

Equation:

$$\begin{array}{rcl} r_1 & = & \int_0^T r_t \alpha\cos\left(2\pi f_c t + \hat{\varphi}\right)\,\mathrm{d}\,t \\ & = & \pm\left(\int_0^T \alpha A\cos(2\pi f_c t + \varphi)\cos\left(2\pi f_c t + \hat{\varphi}\right)\,\mathrm{d}\,t\right) + \int_0^T \alpha\cos\left(2\pi f_c t + \hat{\varphi}\right)N_t\,\mathrm{d}\,t \end{array}$$

Equation:

$$r_1 = \pm\left(\frac{\alpha A}{2}\int_0^T \cos\left(4\pi f_c t + \varphi + \hat{\varphi}\right) + \cos\left(\varphi - \hat{\varphi}\right)\,\mathrm{d}\,t\right) + \eta_1$$

Equation:

$$r_1 = \pm\left(\frac{\alpha A}{2}T\cos\left(\varphi - \hat{\varphi}\right)\right) + \int_0^T \pm\left(\frac{\alpha A}{2}\cos\left(4\pi f_c t + \varphi + \hat{\varphi}\right)\right)\,\mathrm{d}\,t + \eta_1 \pm\left(\frac{\alpha AT}{2}\cos\left(\varphi - \hat{\varphi}\right)\right) + \eta_1$$

where $\eta_1 = \alpha\int_0^T N_t\cos\left(\omega_c t + \hat{\varphi}\right)\,\mathrm{d}\,t$ is zero mean Gaussian with variance $\simeq \frac{\alpha^2 N_0 T}{4}$.

Therefore,

Equation:

$$\begin{array}{rcl} \overline{P_e} & = & Q\left(\frac{2\frac{\alpha AT}{2}\cos\left(\varphi - \hat{\varphi}\right)}{2\sqrt{\frac{\alpha^2 N_0 T}{4}}}\right) \\ & = & Q\left(\cos\left(\varphi - \hat{\varphi}\right)A\sqrt{\frac{T}{N_0}}\right) \end{array}$$

which is not a function of α and depends strongly on phase accuracy.

Equation:

$$P_e = Q\left(\cos\left(\varphi - \hat{\varphi}\right)\sqrt{\frac{2E_s}{N_0}}\right)$$

The above result implies that the amplitude of the local oscillator in the correlator structure does not play a role in the performance of the correlation receiver. However, the accuracy of the phase does indeed play a major role. This point can be seen in the following example:

Example:
Equation:

$$x_{t'} = -1^i A\cos\left(-\left(2\pi f_c t'\right) + 2\pi f_c \tau\right)$$

Equation:

$$x_t = -1^i A\cos\left(2\pi f_c t - \left(2\pi f_c \tau' - 2\pi f_c \tau + \theta'\right)\right)$$

Local oscillator should match to phase θ.

Frequency Shift Keying (FSK)

The data is impressed upon the carrier frequency. Therefore, the M different signals are
Equation:

$$s_m(t) = AP_T(t)\cos(2\pi f_c t + 2\pi\,(m-1)\Delta(f)t + \theta_m)$$

for $m \in \{1, 2, \ldots, M\}$

The M different signals have M different carrier frequencies with possibly different phase angles since the generators of these carrier signals may be different. The carriers are
Equation:

$$f_1 = f_c$$

$$f_2 = f_c + \Delta(f)$$

$$f_M = f_c - M\Delta(f)$$

Thus, the M signals may be designed to be orthogonal to each other.
Equation:

$$\begin{array}{rcl}\langle s_m, s_n\rangle & = & \int_0^T A^2\cos(2\pi f_c t + 2\pi\,(m-1)\Delta(f)t + \theta_m)\cos(2\pi f_c t + 2\pi\,(n-1)\Delta(f)t + \theta_n)\ \mathrm{d}\,t \\ & = & \frac{A^2}{2}\int_0^T \cos(4\pi f_c t + 2\pi\,(n+m-2)\Delta(f)t + \theta_m + \theta_n)\ \mathrm{d}\,t + \frac{A^2}{2}\int_0^T \cos(2\pi\,(m-n)\Delta(f)t + \theta_m \\ & = & \frac{A^2}{2}\frac{\sin(4\pi f_c T + 2\pi(n+m-2)\Delta(f)T + \theta_m + \theta_n) - \sin(\theta_m + \theta_n)}{4\pi f_c + 2\pi(n+m-2)\Delta(f)} + \frac{A^2}{2}\left(\frac{\sin(2\pi(m-n)\Delta(f)T + \theta_m - \theta_n)}{2\pi(m-n)\Delta(f)} - \frac{\sin(\theta_m - \theta_n)}{2\pi(m-n)\Delta(f)}\right)\end{array}$$

If $2f_c T + (n+m-2)\Delta(f)T$ is an integer, and if $(m-n)\Delta(f)T$ is also an integer, then $\langle S_m, S_n\rangle = 0$ if $\Delta(f)T$ is an integer, then $\langle s_m, s_n\rangle \simeq 0$ when f_c is much larger than $\frac{1}{T}$.

In case $\forall m, \theta_m = 0 : (\theta_m = 0)$
Equation:

$$\langle s_m, s_n\rangle \simeq \frac{A^2 T}{2}\,\mathrm{sinc}\,(2\,(m-n)\Delta(f)T)$$

Therefore, the frequency spacing could be as small as $\Delta(f) = \frac{1}{2T}$ since $\mathrm{sinc}\,(x) = 0$ if $x = \pm(1)$ or $\pm(2)$.

If the signals are designed to be orthogonal then the average probability of error for binary FSK with optimum receiver is
Equation:

$$\bar{P}_e = Q\left(\sqrt{\frac{E_s}{N_0}}\right)$$

in AWGN.

Note that $\mathrm{sinc}\,(x)$ takes its minimum value not at $x = \pm(1)$ but at $\pm(1.4)$ and the minimum value is -0.216. Therefore if $\Delta(f) = \frac{0.7}{T}$ then

Equation:

$$\bar{P}_e = Q\left(\sqrt{\frac{1.216E_s}{N_0}}\right)$$

which is a gain of $10 \times \log 1.216 \simeq 0.85d\theta$ over orthogonal FSK.

Differential Phase Shift Keying

The phase lock loop provides estimates of the phase of the incoming modulated signal. A phase ambiguity of exactly π is a common occurance in many phase lock loop (PLL) implementations.

Therefore it is possible that, $\hat{\theta} = \theta + \pi$ without the knowledge of the receiver. Even if there is no noise, if $b = 1$ then $\hat{b} = 0$ and if $b = 0$ then $\hat{b} = 1$.

In the presence of noise, an incorrect decision due to noise may results in a correct final desicion (in binary case, when there is π phase ambiguity with the probability:

Equation:

$$\bar{P_e} = 1 - Q\left(\sqrt{\frac{2E_s}{N_0}}\right)$$

Consider a stream of bits $a_n \in \{0, 1\}$ and BPSK modulated signal

Equation:

$$\sum_n -1^{a_n} A P_T(t - nT)\cos(2\pi f_c t + \theta)$$

In differential PSK, the transmitted bits are first encoded $b_n = a_n \oplus b_{n-1}$ with initial symbol (e.g. b_0) chosen without loss of generality to be either 0 or 1.

Transmitted DPSK signals

Equation:

$$\sum_n -1^{b_n} A P_T(t - nT)\cos(2\pi f_c t + \theta)$$

The decoder can be constructed as

Equation:

$$\begin{aligned} b_{n-1} \oplus b_n &= b_{n-1} \oplus a_n \oplus b_{n-1} \\ &= 0 \oplus a_n \\ &= a_n \end{aligned}$$

If two consecutive bits are detected correctly, if $\hat{b}_n = b_n$ and $\hat{b}_{n-1} = b_{n-1}$ then

Equation:

$$\begin{aligned} \hat{a}_n &= \hat{b}_n \oplus \hat{b}_{n-1} \\ &= b_n \oplus b_{n-1} \\ &= a_n \oplus b_{n-1} \oplus b_{n-1} \\ &= a_n \end{aligned}$$

if $\hat{b}_n = b_n \oplus 1$ and $\hat{b}_{n-1} = b_{n-1} \oplus 1$. That is, two consecutive bits are detected incorrectly. Then,

Equation:

$$\begin{aligned} \hat{a}_n &= \hat{b}_n \oplus \hat{b}_{n-1} \\ &= b_n \oplus 1 \oplus b_{n-1} \oplus 1 \\ &= b_n \oplus b_{n-1} \oplus 1 \oplus 1 \\ &= b_n \oplus b_{n-1} \oplus 0 \\ &= b_n \oplus b_{n-1} \\ &= a_n \end{aligned}$$

If $\hat{b}_n = b_n \oplus 1$ and $\hat{b}_{n-1} = b_{n-1}$, that is, one of two consecutive bits is detected in error. In this case there will be an error and the probability of that error for DPSK is

Equation:

$$\begin{aligned}\bar{P}_e &= \Pr\left[\hat{a}_n \neq a_n\right] \\ &= \Pr\left[\hat{b}_n = b_n, \hat{b}_{n-1} \neq b_{n-1}\right] + \Pr\left[\hat{b}_n \neq b_n, \hat{b}_{n-1} = b_{n-1}\right] \\ &= 2Q\left(\sqrt{\frac{2E_s}{N_0}}\right)\left[1 - Q\left(\sqrt{\frac{2E_s}{N_0}}\right)\right] \simeq 2Q\left(\sqrt{\frac{2E_s}{N_0}}\right)\end{aligned}$$

This approximation holds if Q is small.

Digital Transmission over Baseband Channels

Until this point, we have considered data transmissions over simple additive Gaussian channels that are not time or band limited. In this module we will consider channels that do have bandwidth constraints, and are limited to frequency range around zero (DC). The channel is best modified as $g(t)$ is the impulse response of the baseband channel.

Consider modulated signals $x_t = s_m(t)$ for $0 \leq t \leq T$ for some $m \in \{1, 2, \ldots, M\}$. The channel output is then

Equation:

$$\begin{aligned} r_t &= {}_{-\infty}^{\infty} x_\tau g(t-\tau)\ \mathrm{d}\ \tau + N_t \\ &= {}_{-\infty}^{\infty} S_m(\tau) g(t-\tau)\ \mathrm{d}\ \tau + N_t \end{aligned}$$

The signal contribution in the frequency domain is

Equation:

$$\forall f : \quad S_m(f) = S_m(f) G(f)$$

The optimum matched filter should match to the filtered signal:

Equation:

$$\forall f : \quad H_m^{\mathrm{opt}}(f) = S_m(f) G(f) e^{(-i) 2\pi f t}$$

This filter is indeed **optimum** (i.e., it maximizes signal-to-noise ratio); however, it requires knowledge of the channel impulse response. The signal energy is changed to

Equation:

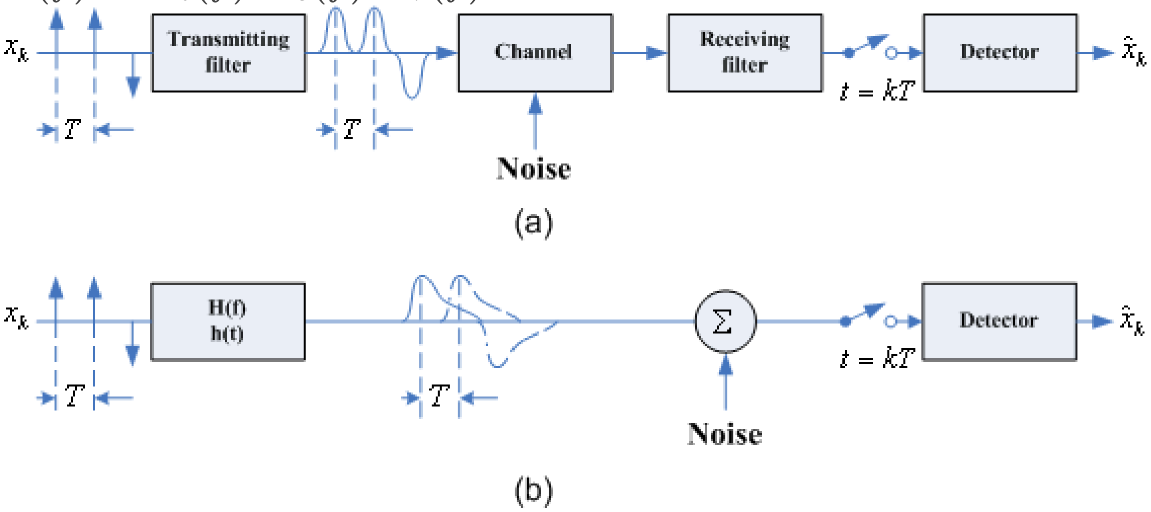
$$E_{\tilde{s}} = {}_{-\infty}^{\infty} \quad S_m(f) \quad {}^2\ \mathrm{d}\ f$$

The band limited nature of the channel and the stream of time limited modulated signal create aliasing which is referred to as **intersymbol interference**. We will investigate ISI for a general PAM signaling.

Introduction to ISI

A typical baseband digital system is described in Figure 1(a). At the transmitter, the modulated pulses are filtered to comply with some bandwidth constraint. These pulses are distorted by the reactances of the cable or by fading in the wireless systems. Figure 1(b) illustrates a convenient model, lumping all the filtering into one overall equivalent system transfer function.

$$H(f) = H_t(f).H_c(f).H_r(f)$$

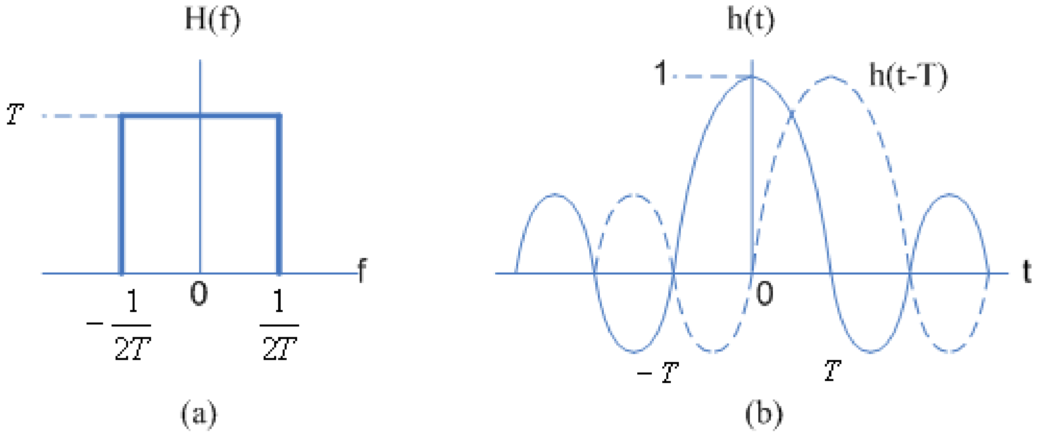


Intersymbol interference in the detection process. (a) Typical baseband digital system. (b) Equivalent model

Due to the effects of system filtering, the received pulses can overlap one another as shown in Figure 1(b). Such interference is termed InterSymbol Interfernce (ISI). Even in the absence of noise, the effects of filtering and channel-induced distortion lead to ISI.

Nyquist investigated and showed that theoretical minimum system bandwidth needed in order to detect R_s symbols/s, without ISI, is $R_s/2$ or $1/2\mathrm{T}$ hertz. For baseband systems, when $H(f)$ is such a filter with single-sided bandwidth $1/2\mathrm{T}$ (the ideal Nyquist filter) as shown in figure 2a, its impulse response is of the form $h(t) = \mathrm{sinc}(t/T)$, shown in figure 2b. This

$\mathrm{sin}c(t/T)$-shaped pulse is called the ideal Nyquist pulse. Even though two successive pulses $h(t)$ and $h(t-T)$ with long tail, the figure shows all tail of $h(t)$ passing through zero amplitude at the instant when $h(t-T)$ is to be sampled. Therefore, assuming that the synchronization is perfect, there will be no ISI.



Nyquist channels for zero ISI. (a) Rectangular system transfer function H(f). (b) Received pulse shape $h(t) = \mathrm{sinc}(t/T)$

Figure 2 Nyquist channels for zero ISI. (a) Rectangular system transfer function H(f). (b) Received pulse shape $h(t) = \mathrm{sinc}(t/T)$

The names "Nyquist filter" and "Nyquist pulse" are often used to describe the general class of filtering and pulse-shaping that satisfy zero ISI at the sampling points. Among the class of Nyquist filters, the most popular ones are the raised cosine and root-raised cosine.

A fundamental parameter for communication system is bandwidth efficiency, R/W bits/s/Hz. For ideal Nyquist filtering, the theoretical maximum symbol-rate packing without ISI is $2\text{symbols}/s/\text{Hz}$. For example, with 64-ary PAM, $M = 64 = 2^6$ amplitudes, the theoretical maximum bandwidth efficiency is possible without ISI is $6\text{bits}/\text{symbol}.2\text{symbols}/s/\text{Hz} = 12\text{bits}/s/\text{Hz}$.

Pulse Amplitude Modulation Through Bandlimited Channel

Consider a PAM system b_{-10},…, b_{-1}, b_0 b_1,…

This implies
Equation:

$$\forall a_n, a_n \in \{\text{M levels of amplitude}\} : \left(x_t = \sum_{n=-\infty}^{\infty} a_n s(t - nT)\right)$$

The received signal is
Equation:

$$\begin{array}{rcl} r_t & = & \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_n s(t-(\tau - nT))g(\tau)\ \mathrm{d}\ \tau + N_t \\ & = & \sum_{n=-\infty}^{\infty} a_n \int_{-\infty}^{\infty} s(t-(\tau - nT))g(\tau)\ \mathrm{d}\ \tau + N_t \\ & = & \sum_{n=-\infty}^{\infty} a_n \tilde{s}(t - nT) + N_t \end{array}$$

Since the signals span a one-dimensional space, one filter matched to $\tilde{s}(t) = \bar{s}g(t)$ is sufficient.

The matched filter's impulse response is
Equation:

$$\forall t : \left(h^{\mathrm{opt}}(t) = \bar{s}g(T - t)\right)$$

The matched filter output is
Equation:

$$\begin{array}{rcl} y(t) & = & \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_n \tilde{s}(t-(\tau - nT))h^{\mathrm{opt}}(\tau)\ \mathrm{d}\ \tau + \nu(t) \\ & = & \sum_{n=-\infty}^{\infty} a_n \int_{-\infty}^{\infty} \tilde{s}(t-(\tau - nT))h^{\mathrm{opt}}(\tau)\ \mathrm{d}\ \tau + \nu(t) \\ & = & \sum_{n=-\infty}^{\infty} a_n u(t - nT) + \nu(t) \end{array}$$

The decision on the k^{th} symbol is obtained by sampling the MF output at kT:

Equation:

$$y(kT) = \sum_{n=-\infty}^{\infty} a_n u(kT - nT) + \nu(kT)$$

The k^{th} symbol is of interest:

Equation:

$$y(kT) = a_k u(0) + \sum_{n=-\infty}^{\infty} a_n u(kT - nT) + \nu(kT)$$

where $n \neq k$.

Since the channel is bandlimited, it provides memory for the transmission system. The effect of old symbols (possibly even future signals) lingers and affects the performance of the receiver. The effect of ISI can be eliminated or controlled by proper design of **modulation signals** or **precoding** filters at the transmitter, or by **equalizers** or **sequence detectors** at the receiver.

Precoding and Bandlimited Signals

Precoding

The data symbols are manipulated such that

Equation:

$$y_k(kT) = a_k u(0) + \mathrm{ISI} + \nu(kT)$$

Design of Bandlimited Modulation Signals

Recall that modulation signals are

Equation:

$$X_t = \sum_{n=-\infty}^{\infty} a_n s(t - nT)$$

We can design $s(t)$ such that

Equation:

$$u(nT) = \begin{array}{l} \text{large} \;\text{ if }\; n = 0 \\ \text{zero or small }\;\text{if }\; n \neq 0 \end{array}$$

where $y(kT) = a_k u(0) + {}_{n=-\infty}^{\infty} a_n u(kT - nT) + \nu(kT)$ (ISI is the sum term, and once again, $n \neq k$.) Also, $y(nT) = sgh^{\mathrm{opt}}(nT)$ The signal $s(t)$ can be designed to have reduced ISI.

Design Equalizers at the Receiver

Linear equalizers or decision feedback equalizers reduce ISI in the statistic y_t

Maximum Likelihood Sequence Detection

Equation:

$$y(kT) = \sum_{n=-\infty}^{\infty} a_n\,(kT - nT) + \nu(k(T))$$

By observing $y(T)$,$y(2T)$,. . . the date symbols are observed frequently. Therefore, ISI can be viewed as diversity to increase performance.

Pulse Shaping to Reduce ISI

The Raised-Cosine Filter

Transfer function beloging to the Nyquist class (zero ISI at the sampling time) is called the raised-cosine filter. It can be express as

$$H(f) = \begin{array}{ll} 1 & \mid f \mid < 2\mathrm{W}_0 - W \\ \cos^2\left(\frac{\pi}{4}\frac{|f|+W-2\mathrm{W}_0}{W-W_0}\right) & 2\mathrm{W}_0 - W < \mid f \mid < W(1\mathrm{a}) \\ 0 & \mid f > W \mid \end{array}$$

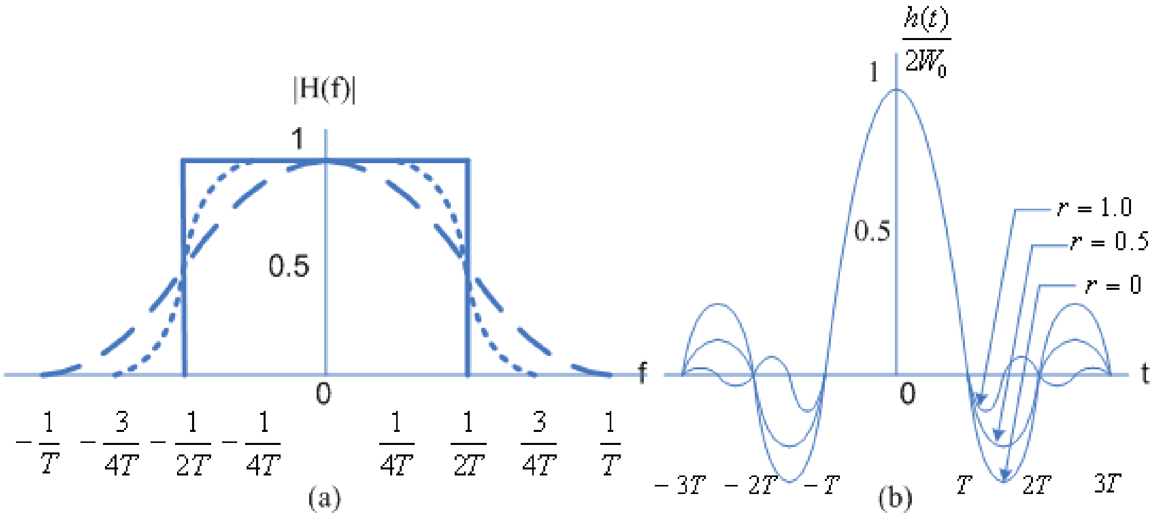
$$h(t) = 2\mathrm{W}_0\mathrm{sinc}(2\mathrm{W}_0 t)\frac{\cos[2\pi(W-W_0)t]}{1-[4(W-W_0)t]^2} \text{ (1b)}$$

Where W is the absolute bandwidth. $W_0 = 1/2\mathrm{T}$ represent the minimum bandwidth for the rectangular spectrum and the -6 dB bandwith (or half-amplitude point) for the raised-cosine spectrum. $W - W_0$ is termed the "excess bandwith"

The roll-off factor is defined to be $r = \frac{W-W_0}{W_0}$ (2), where $0 \leq r \leq 1$

With the Nyquist constrain $W_0 = R_s/2$ equation (2) can be rewriten as

$$W = \frac{1}{2}(1+r)R_s$$



Raised-cosine filter characteristics. (a) System transfer function.

(b) System impulse response

The raised-cosine characteristic is illustrate in figure 1 for $r = 0$,$r = 0.5$,$r = 1$. When $r = 1$, the required excess bandwidth is 100 %, and the system can provide a symbol rate of R_s symbols/s using a bandwidth of R_s herts (twice the Nyquist minimum bandwidth), thus yielding asymbol-rate packing 1 symbols/s/Hz.

The lager the filter roll-off, the shorter will be the pulse tail. Small tails exhibit less sensitivity to timing errors and thus make for small degradation due to ISI.

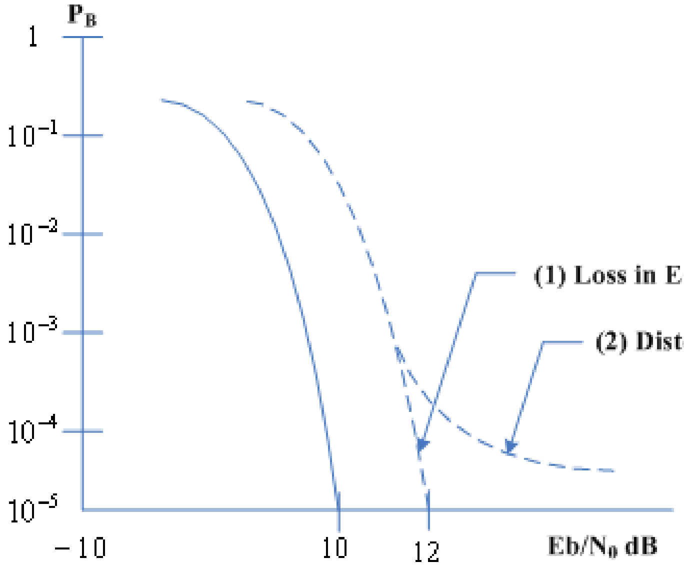
The smaller the filter roll-off the smaller will be the excess bandwidth. The cost is longer pulse tails, larger pulse amplitudes, and thus, greater sensitivity to timing errors.

The Root Raised-Cosine Filter

Recall that the raised-cosine frequency transfer function describes the composite $H(f)$ including transmitting filter, channel filter and receiving filter. The filtering at the receiver is chosen so that the overall transfer function is a form of raised-cosine. Often this is accomplished by choosing both the receiving filter and the transmitting filter so that each has a transfer function known as a root raised cosine. Neglecting any channel-induced ISI, the product of these root-raised cosine functions yields the composite raised-cosine system transfer function.

Two Types of Error-Performance Degradation

Error-performance degradation can be classifyed in two group. The first one is due to a decrease in received signal power or an increase in noise or inteference power, giving rise to a loss in signal-to-noise ratio E_B/N_0. The second one is due to signal distortion such as ISI.

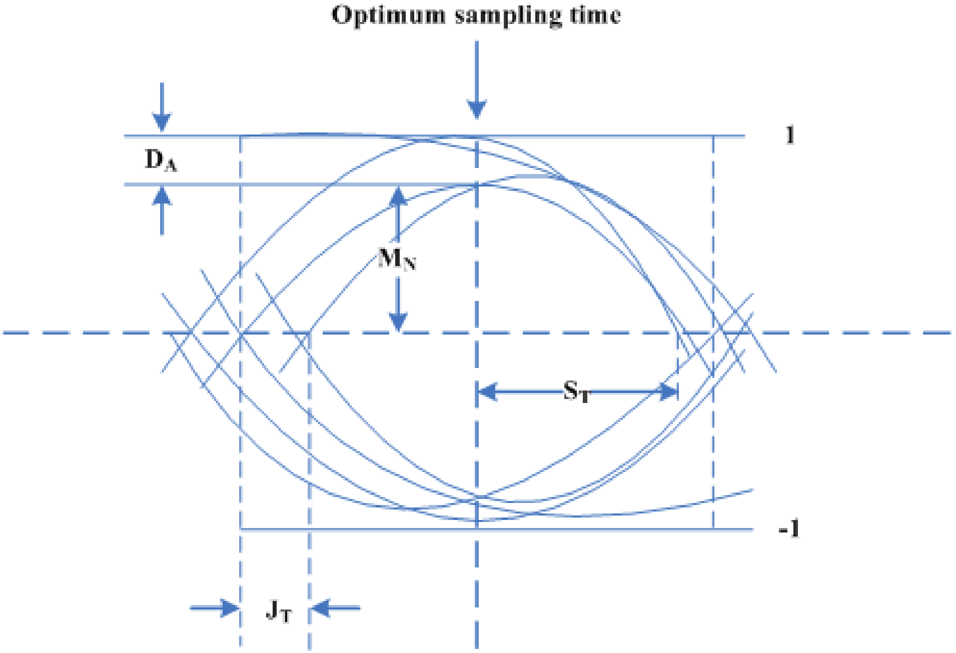


Bit error probability

Suppose that we need a communication system with a bit-error probability P_B versus E_b/N_0 characteristic corresponding to the solid-line curve plotted in figure 1. Suppose that after the system is configured, the performance dose not follow the theoretical curve, but in facts follows the dashed line plot (1). A loss in E_b/N_0 due to some signal losses or an increased level of noise or interference. This loss in E_B/N_0 is not so terrible when compared with possible effects of degradation caused by a distortion mechanism corresponding to the dashed line plot (2). Instead of suffering a simple loss in signal-to-noise ratio there is a degradation effect brought about by ISI. If there is no solution to this problem, there is no a mount of E_B/N_0 that will improve this problem. More E_B/N_0 can not help the ISI problem because a incresing in E_B/N_0 dose not make change in overlapped pulses.

Eye Pattern

An eye pattern is the display that results from measuring a system' s response to baseband signals in a prescribed way.



Eye pattern

Figure 1 describe the eye pattern that results for binary binary pulse signalling. The width of the opening indicates the time over which sampling for detection might be performed. The optimum sampling time corresponds to the maxmum eye opening, yielding the greatest protection against noise. If there were no filtering in the system then the system would look like a box rather than an eye. In figure 1, D_A, the range of amplitude differences of the zero crossings, is a measure of distortion caused by ISI.

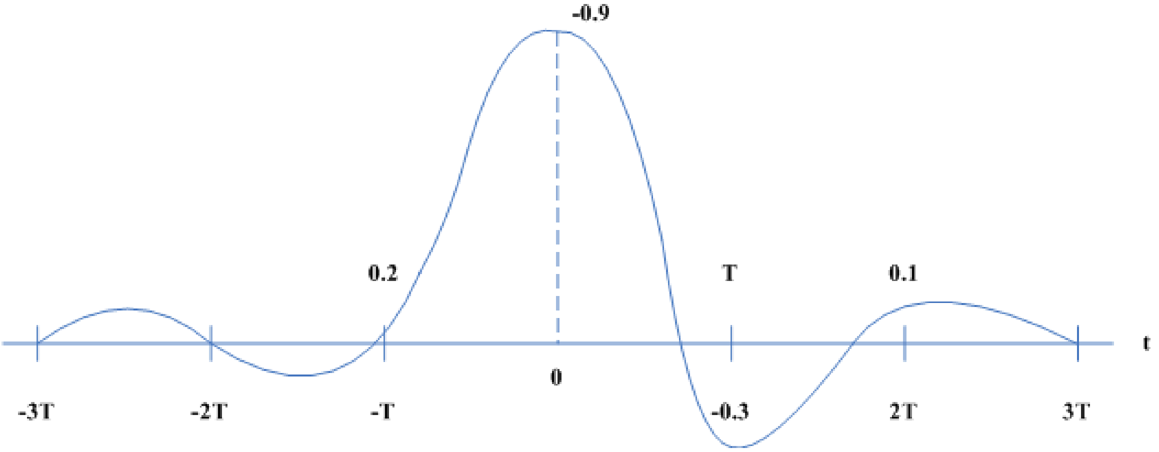
J_T, the range of amplitude differences of the zero crossing , is a measure of the timmung jitter. M_N is a measure of noise margin. S_T is mesuare of sensity-to-timing error.

In general, the most frequent use of the eye pattern is for qualitatively assessing the extent of the ISI. As the eye closes, ISI is increase; as the eye opens, ISI is decreaseing.

Transversal Equalizer

A training sequence used for equalization is often chosen to be a noise-like sequence which is needed to estimate the channel frequency response.

In the simplest sense, training sequence might be a single narrow pulse, but a pseudonoise (PN) signal is preferred in practise because the PN signal has larger average power and hence larger SNR for the same peak transmitted power.

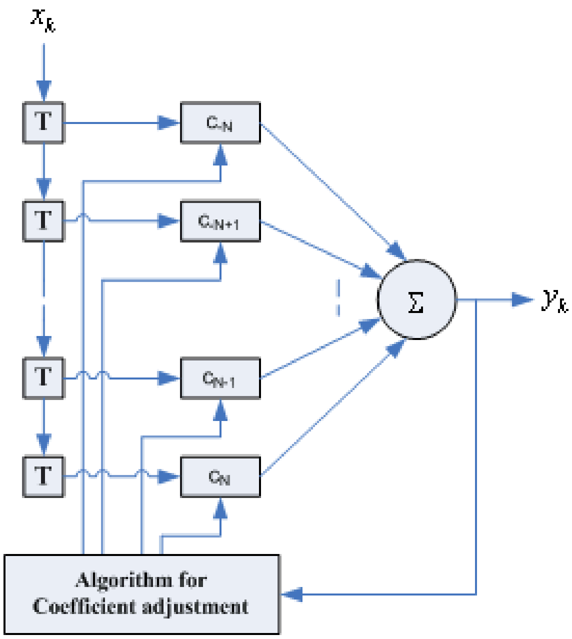


Received pulse exhibiting distortion

Consider that a single pulse was transmitted over a system designated to have a raised-cosine transfer function $H_{\mathrm{RC}}(t) = H_t(f).H_r(f)$, also consider that the channel induces ISI, so that the received demodulated pulse exhibits distortion, as shown in figure 1, such that the pulse sidelobes do not go through zero at sample times. To achieve the desired raised-cosine transfer function, the equalizing filter should have a frequency response

$H_e(f) = \frac{1}{H_c(f)} = \frac{1}{|H_c(f)|} e^{-j\theta_c(f)}$ (1)

In other words, we would like the equalizing filter to generate a set of canceling echoes. The transversal filter, illustrated in figure 2, is the most popular form of an easily adjustable equalizing filter consisting of a delay line with T-second taps (where T is the symbol duration). The tab weights could be chosen to force the system impulse response to zero at all but one of the sampling times, thus making $H_e(f)$ correspond exactly to the inverse of the channel transfer function $H_c(f)$



Transversal filter

Consider that there are 2N $+ 1$ taps with weights c_{-N},c_{-N+1},...c_N . Output samples $z(k)$ are the convolution the input sample $x(k)$ and tap weights c_n as follows:

$$z(k) = \sum_{n=-N}^{N} x(k-n)c_n k = -2\text{N},...2\text{N}(2)$$

By defining the vectors z and c and the matrix x as respectively,

$$z = \begin{array}{c} z(-2\text{N}) \\ \vdots \\ z(0) \\ \vdots \\ z(2\text{N}) \end{array} \quad c = \begin{array}{c} c_{-N}) \\ \vdots \\ c_0 \\ \vdots \\ c_N \end{array}$$

$$x = \begin{array}{cccccc} x(-N) & 0 & 0 & \dots & 0 & 0 \\ x(-N+1) & x(-N) & 0 & \dots & \dots & \dots \\ \vdots & & & \vdots & & \vdots \\ x(N) & x(N-1) & x(N-2) & \dots & x(-N+1) & x(-N) \\ \vdots & & & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & x(N) & x(N-1) \\ 0 & 0 & 0 & \dots & 0 & x(N) \end{array}$$

We can describe the relationship among $z(k)$, $x(k)$ and c_n more compactly as

$z = x.c$(3a)

Whenever the matrix x is square, we can find c by solving the following equation:

$c = x^{-1}z$(3b)

Notice that the index k was arbitrarily chosen to allow for 4N + 1 sample points. The vectors z and c have dimensions 4N + 1 and 2N + 1. Such equations are referred to as an overdetermined set. This problem can be solved in deterministic way known as the zero-forcing solution, or, in a statistical way, known as the minimum mean-square error (MSE) solution.

Zero-Forcing Solution

At first, by disposing top N rows and bottom N rows, matrix x is transformed into a square matrix of dimension 2N + 1 by 2N + 1. Then equation $c = x^{-1}z$ is used to solve the $2\mathrm{N} + 1$ simultaneous equations for the set of $2\mathrm{N} + 1$ weights c_n. This solution minimizes the peak ISI distortion by selecting the C_n weight so that the equalizer output is forced to zero at N sample points on either side of the desired pulse.

$$z(k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \pm 3 \end{cases} (4)$$

For such an equalizer with finite length, the peak distortion is guaranteed to be minimized only if the eye pattern is initially open. However, for high-speed transmission and channels introducing much ISI, the eye is often closed before equalization. Since the zero-forcing equalizer neglects the effect of noise, it is not always the best system solution.

Minimum MSE Solution

A more robust equalizer is obtained if the c_n tap weights are chose to minimize the mean-square error (MSE) of all the ISI term plus the noise power at the out put of the equalizer. MSE is defined as the expected value of the squared difference between the desire data symbol and the estimated data symbol.

By multiplying both sides of equation (4) by x^T, we have

$$x^T z = x^T \mathrm{xc}(5)$$

And

$$R_{\mathrm{xz}} = R_{\mathrm{xx}} c \ (6)$$

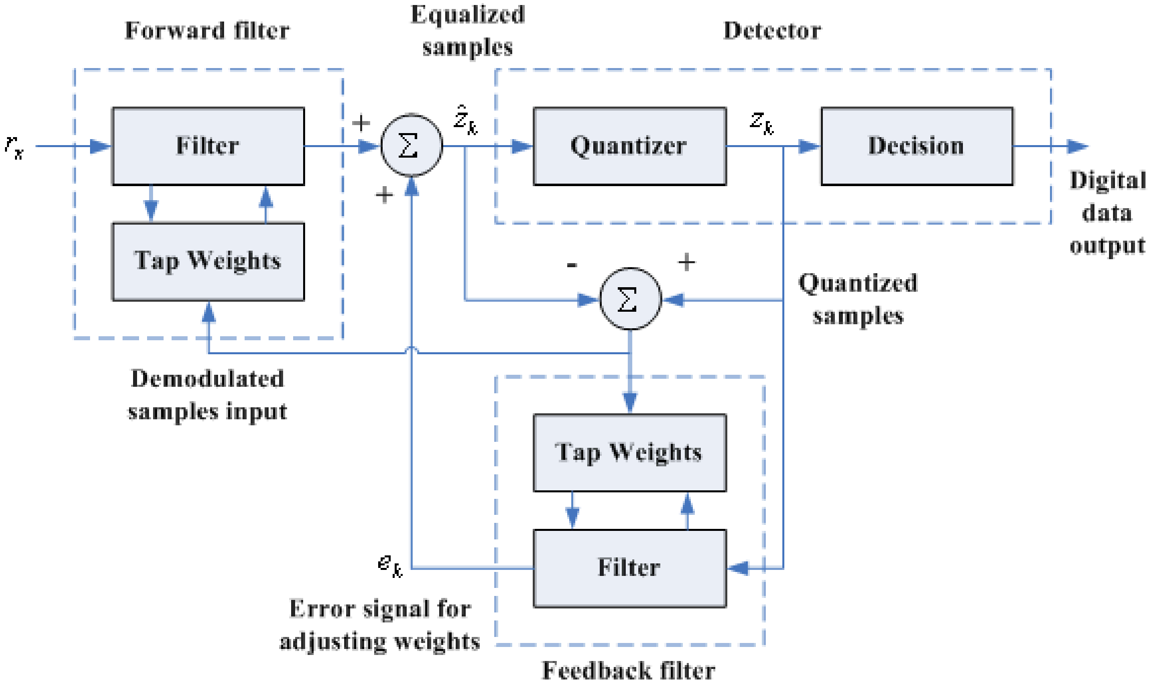
Where $R_{\mathrm{xz}} = x^T z$ is called the cross-correlation vector and $R_{\mathrm{xx}} = x^T x$ is call the autocorrelation matrix of the input noisy signal. In practice, R_{xz} and R_{xx} are unknown, but they can be approximated by transmitting a test signal and using time average estimated to solve for the tap weights from equation (6) as follows:

$$c = R_{\mathrm{xx}}^{-1} R_{\mathrm{xz}}$$

Most high-speed telephone-line modems use an MSE weight criterion because it is superior to a zero-forcing criterion; it is more robust in the presence of noise and large ISI.

Decision Feedback Equalizer

The basic limitation of a linear equalizer, such as the transversal filter, is the poor perform on channel having spectral nulls. A decision feedback equalizer (DFE) is a nonlinear equalizer that uses previous detector decision to eliminate the ISI on pulses that are currently being demodulated. In other words, the distortion on a current pulse that was caused by previous pulses is subtracted.



Decision feedback Equalizer

Figure 1 shows a simplified block diagram of a DFE where the forward filter and the feedback filter can each be a linear filter, such as transversal filter. The nonlinearity of the DFE stems from the nonlinear characteristic of the detector that provides an input to the feedback filter. The basic idea of a DFE is that if the values of the symbols previously detected are known, then ISI contributed by these symbols can be canceled out exactly at the output of the forward filter by subtracting past symbol values with appropriate weighting. The forward and feedback tap weights can be adjusted simultaneously to fulfill a criterion such as minimizing the MSE.

The advantage of a DFE implementation is the feedback filter, which is additionally working to remove ISI, operates on noiseless quantized levels, and thus its output is free of channel noise.

Adaptive Equalization

Another type of equalization, capable of tracking a slowly time-varying channel response, is known as adaptive equalization. It can be implemented to perform tap-weight adjustments periodically or continually. Periodic adjustments are accomplished by periodically transmitting a preamble or short training sequence of digital data known by the receiver. Continual adjustment are accomplished by replacing the known training sequence with a sequence of data symbols estimated from the equalizer output and treated as known data. When performed continually and automatically in this way, the adaptive procedure is referred to as decision directed.

If the probability of error exceeds one percent, the decision directed equalizer might not converge. A common solution to this problem is to initialize the equalizer with an alternate process, such as a preamble to provide good channel-error performance, and then switch to decision-directed mode.

The simultaneous equations described in equation (3) of module “Transversal Equalizer”, do not include the effects of channel noise. To obtain stable solution to the filter weights, it is necessary that the data be averaged to obtain the stable signal statistic, or the noisy solution obtained from the noisy data must be averaged. The most robust algorithm that average noisy solution is the least-mean-square (LMS) algorithm. Each iteration of this algorithm uses a noisy estimate of the error gradient to adjust the weights in the direction to reduce the average mean-square error.

The noisy gradient is simply the product $e(k) r_x$ of an error scalar $e(k)$and the data vector r_x.

$e(k) = z(k) - \hat{z}(k)$ (1)

Where $z(k)$ and $\hat{z}(k)$ are the desired output signal (a sample free of ISI) and the estimate at time k.

$\hat{z}(k) = c^T r_x = \sum_{n=-N}^{N} x(k-n) c_n$ (2)

Where c^T is the transpose of the weight vector at time k.

Iterative process that updates the set of weights is obtained as follows:

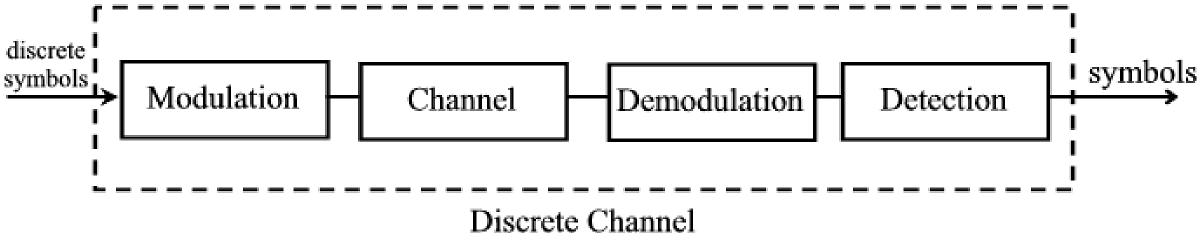
$c(k+1) = c(k) + \mathit{\Delta} e(k) r_x$ (3)

Where $c(k)$ is the vector of filter weights at time k, and $\mathit{\Delta}$ is a small term that limits the coefficient step size and thus controls the rate of convergence of the algorithm as well as the variance of the steady state solution. Stability is assured if the parameter $\mathit{\Delta}$ is smaller than the reciprocal of the energy of the data in the filter. Thus, while we want the convergence parameter $\mathit{\Delta}$ to be large for fast convergence but not so large as to be unstable, we also want it to be small enough for low variance.

Channel Capacity

In the previous section, we discussed information sources and quantified information. We also discussed how to represent (and compress) information sources in binary symbols in an efficient manner. In this section, we consider channels and will find out how much information can be sent through the channel reliably.

We will first consider simple channels where the input is a discrete random variable and the output is also a discrete random variable. These discrete channels could represent analog channels with modulation and demodulation and detection.



Let us denote the input sequence to the channel as

Equation:

$$\boldsymbol{X} = \begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{matrix}$$

where $X_i \in X$ a discrete symbol set or input alphabet.

The channel output

Equation:

$$\boldsymbol{Y} = \begin{matrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{matrix}$$

where $Y_i \in Y$ a discrete symbol set or output alphabet.

The statistical properties of a channel are determined if one finds $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ for all $\boldsymbol{y} \in Y^n$ and for all $\boldsymbol{x} \in X^n$. A discrete channel is called a **discrete memoryless channel** if

Equation:

$$p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p_{Y_i|X_i}(y_i|x_i)$$

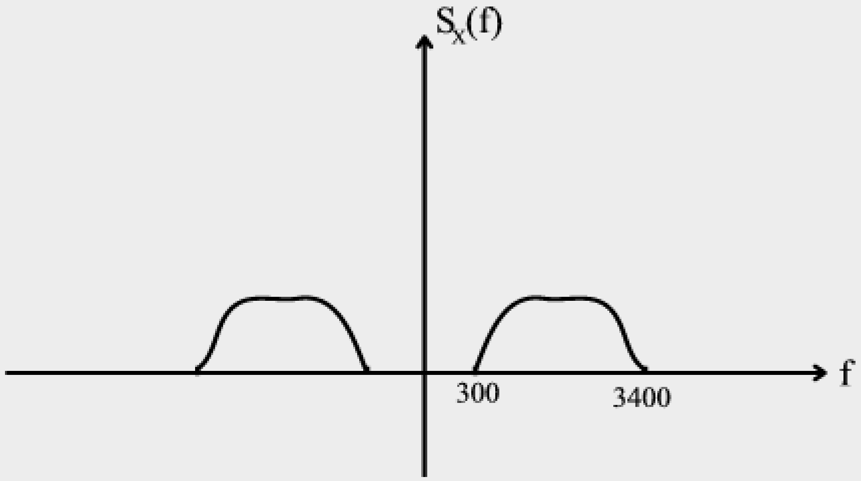
for all $\boldsymbol{y} \in Y^n$ and for all $\boldsymbol{x} \in X^n$.

Example:

A binary symmetric channel (BSC) is a discrete memoryless channel with binary input and binary output and $p_{Y|X}(\mathrm{y}=0|\mathrm{x}=1) = p_{Y|X}(\mathrm{y}=1|\mathrm{x}=0)$. As an example, a white Gaussian channel with antipodal signaling and matched filter receiver has probability of error of $Q\sqrt{\frac{2E_s}{N_0}}$. Since the error is symmetric with respect to the transmitted bit, then

Equation:

$$\begin{aligned} p_{Y|X}(0|1) &= p_{Y|X}(1|0) \\ &= Q \quad \sqrt{\frac{2E_s}{N_0}} \\ &= \varepsilon \end{aligned}$$



It is interesting to note that every time a BSC is used one bit is sent across the channel with probability of error of ε. The question is how much information or how many bits can be sent per channel use, reliably. Before we consider the above question a few definitions are essential. These are discussed in mutual information.

Mutual Information

Recall that

Equation:

$$H(X,Y) = -\sum_{xx}\sum_{yy} \mathrm{p}_{X,Y}\ (x,y)\log \mathrm{p}_{X,Y}\ (x,y)$$

Equation:

$$H(Y) + H(X|Y) = H(X) + H(Y|X)$$

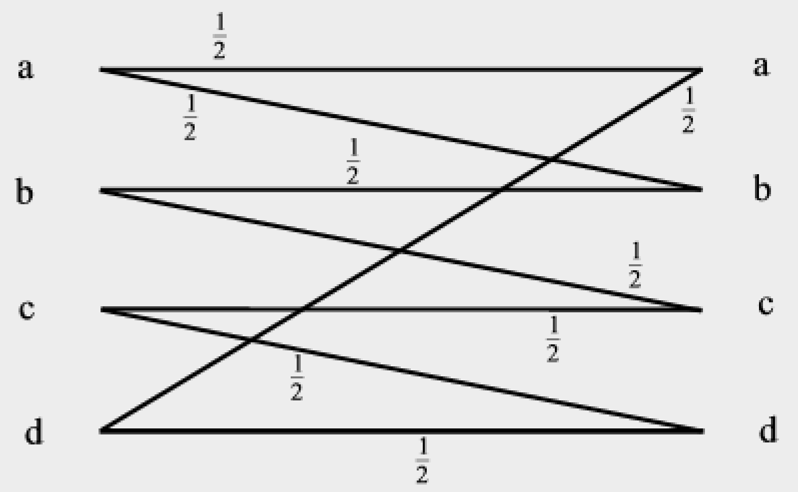
Mutual Information

> The mutual information between two discrete random variables is denoted by $\mathscr{I}(X;Y)$ and defined as
>
> **Equation:**
>
> $$\mathscr{I}(X;Y) = H(X) - H(X|Y)$$
>
> Mutual information is a useful concept to measure the amount of information shared between input and output of noisy channels.

In our previous discussions it became clear that when the channel is noisy there may not be reliable communications. Therefore, the limiting factor could very well be reliability when one considers noisy channels. Claude E. Shannon in 1948 changed this paradigm and stated a theorem that presents the rate (speed of communication) as the limiting factor as opposed to reliability.

Example:

Consider a discrete memoryless channel with four possible inputs and outputs.

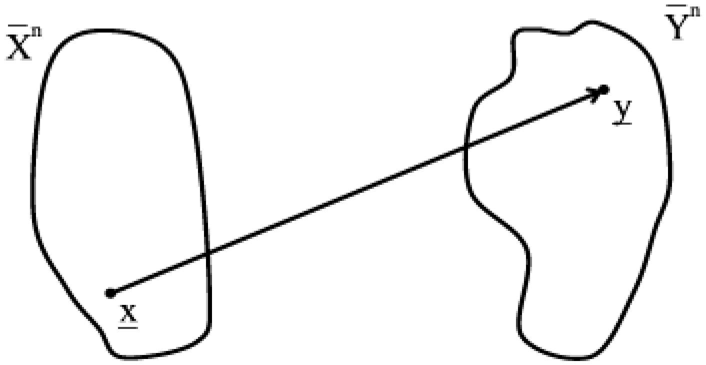


Every time the channel is used, one of the four symbols will be transmitted. Therefore, 2 bits are sent per channel use. The system, however, is very unreliable. For example, if "a" is received, the receiver can not determine, reliably, if "a" was transmitted or "d". However, if the transmitter and receiver agree to only use symbols "a" and "c" and never use "b" and "d", then the transmission will always be reliable, but 1 bit is sent per channel use. Therefore, the rate of transmission was the limiting factor and not reliability.

This is the essence of Shannon's noisy channel coding theorem, i.e., using only those inputs whose corresponding outputs are disjoint (e.g., far apart). The concept is appealing, but does not seem possible with binary channels since the input is either zero or one. It may work if one considers a vector of binary inputs referred to as the extension channel.

$$\mathbf{X\ input\ vector} = \begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{matrix} \quad \in X^n = \{0,1\}^n$$

$$\textbf{Y output vector} = \begin{matrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{matrix} \in Y^n = \{0,1\}^n$$



This module provides a description of the basic information necessary to understand Shannon's Noisy Channel Coding Theorem. However, for additional information on typical sequences, please refer to Typical Sequences.

Typical Sequences

If the binary symmetric channel has crossover probability ε then if $\boldsymbol{x}$ is transmitted then by the Law of Large Numbers the output $\boldsymbol{y}$ is different from $\boldsymbol{x}$ in $n\varepsilon$ places if n is very large.
Equation:

$$d_H(\boldsymbol{x}, \boldsymbol{y}) \simeq n\varepsilon$$

The number of sequences of length n that are different from $\boldsymbol{x}$ of length n at $n\varepsilon$ is
Equation:

$$\binom{n}{n\varepsilon} = \frac{n!}{(n\varepsilon)!\,(n-n\varepsilon)!}$$

Example:
$\boldsymbol{x} = (000)^T$ and $\varepsilon = \frac{1}{3}$ and $n\varepsilon = 3 \times \frac{1}{3}$. The number of output sequences different from $\boldsymbol{x}$ by one element: $\frac{3!}{1!2!} = \frac{3\times2\times1}{1\times2} = 3$ given by $(101)^T$, $(011)^T$, and $(000)^T$.

Using Stirling's approximation
Equation:

$$n! \simeq n^n e^{-n}\sqrt{2\pi n}$$

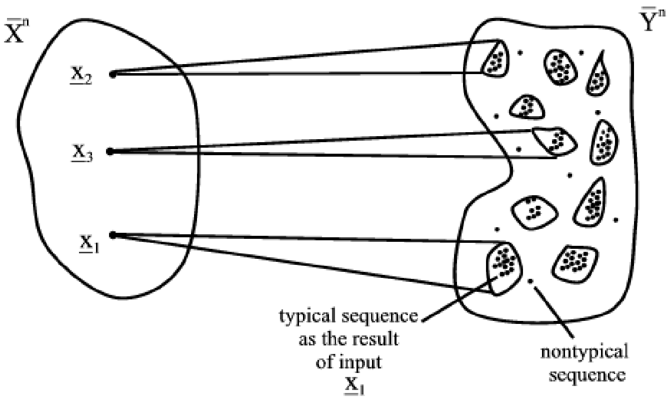
we can approximate
Equation:

$$\binom{n}{n\varepsilon} \simeq 2^{\mathrm{n}((-(\varepsilon\log_2\varepsilon))-(1-\varepsilon)\log_2(1-\varepsilon))} = 2^{nH_b(\varepsilon)}$$

where $H_b(\varepsilon) \equiv (-(\varepsilon\log_2\varepsilon)) - (1-\varepsilon)\log_2(1-\varepsilon)$ is the entropy of a binary memoryless source. For any $\boldsymbol{x}$ there are $2^{nH_b(\varepsilon)}$ highly probable outputs that correspond to this input.

Consider the output vector $\boldsymbol{Y}$ as a very long random vector with entropy $nH(Y)$. As discussed earlier, the number of typical sequences (or highly probably) is roughly $2^{nH(Y)}$. Therefore, 2^n is the total number of binary sequences, $2^{nH(Y)}$ is the number of typical sequences, and $2^{nH_b(\varepsilon)}$ is the number of elements in a group of possible outputs for one input vector. The maximum number of input sequences that produce nonoverlapping output sequences
Equation:

$$\begin{aligned} M &= \frac{2^{nH(Y)}}{2^{nH_b(\varepsilon)}} \\ &= 2^{n(H(Y)-H_b(\varepsilon))} \end{aligned}$$



The number of distinguishable input sequences of length n is
Equation:

$$2^{n(H(Y)-H_b(\varepsilon))}$$

The number of information bits that can be sent across the channel reliably per n channel uses $n\left(H(Y)-H_b(\varepsilon)\right)$ The maximum reliable transmission rate per channel use
Equation:

$$\begin{aligned} R &= \frac{\log_2 M}{n} \\ &= \frac{n(H(Y)-H_b(\varepsilon))}{n} \\ &= H(Y)-H_b(\varepsilon) \end{aligned}$$

The maximum rate can be increased by increasing $H(Y)$. Note that $H_b(\varepsilon)$ is only a function of the crossover probability and can not be minimized any further.

The entropy of the channel output is the entropy of a binary random variable. If the input is chosen to be uniformly distributed with $p_X(0)=p_X(1)=\frac{1}{2}$.

Then
Equation:

$$\begin{aligned} p_Y(0) &= 1p_X(0)+\varepsilon p_X(1) \\ &= \frac{1}{2} \end{aligned}$$

and
Equation:

$$\begin{aligned} p_Y(1) &= 1p_X(1)+\varepsilon p_X(0) \\ &= \frac{1}{2} \end{aligned}$$

Then, $H(Y)$ takes its maximum value of 1. Resulting in a maximum rate $R=1-H_b(\varepsilon)$ when $p_X(0)=p_X(1)=\frac{1}{2}$. This result says that ordinarily one bit is transmitted across a BSC with reliability

$1-\varepsilon$. If one needs to have probability of error to reach zero then one should reduce transmission of information to $1-H_b(\varepsilon)$ and add redundancy.

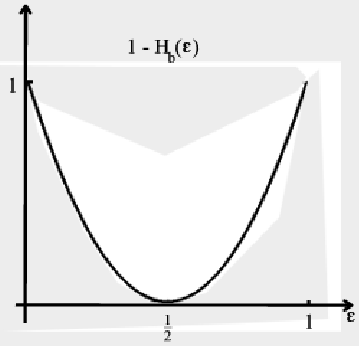
Recall that for Binary Symmetric Channels (BSC)
Equation:

$$\begin{aligned} H(Y|X) &= p_x(0)H(Y|X=0)+p_x(1)H(Y|X=1) \\ &= p_x(0)\left(-\left((1-\varepsilon)\log_2(1-\varepsilon)-\varepsilon\log_2\varepsilon\right)\right)+p_x(1)\left(-\left((1-\varepsilon)\log_2(1-\varepsilon)-\varepsilon\log_2\varepsilon\right)\right) \\ &= \left(-\left((1-\varepsilon)\log_2(1-\varepsilon)\right)\right)-\varepsilon\log_2\varepsilon \\ &= H_b(\varepsilon) \end{aligned}$$

Therefore, the maximum rate indeed was
Equation:

$$\begin{aligned} R &= H(Y)-H(Y|X) \\ &= \mathscr{I}(X;Y) \end{aligned}$$

Example:
The maximum reliable rate for a BSC is $1-H_b(\varepsilon)$. The rate is 1 when $\varepsilon=0$ or $\varepsilon=1$. The rate is 0 when $\varepsilon=\frac{1}{2}$



This module provides background information necessary for an understanding of Shannon's Noisy Channel Coding Theorem. It is also closely related to material presented in Mutual Information.

Shannon's Noisy Channel Coding Theorem

It is highly recommended that the information presented in Mutual Information and in Typical Sequences be reviewed before proceeding with this document. An introductory module on the theorem is available at Noisy Channel Theorems .

Theorem

Shannon's Noisy Channel Coding

The capacity of a discrete-memoryless channel is given by

Equation:

$$C = \max_{p_X(x)} \{\mathscr{I}(X;Y) | p_X(x)\}$$

where $\mathscr{I}(X;Y)$ is the mutual information between the channel input X and the output Y. If the transmission rate R is less than C, then for any $\varepsilon > 0$ there exists a code with block length n large enough whose error probability is less than ε. If $R > C$, the error probability of any code with any block length is bounded away from zero.

Example:

If we have a binary symmetric channel with cross over probability 0.1, then the capacity $C \simeq 0.5$ bits per transmission. Therefore, it is possible to send 0.4 bits per channel through the channel reliably. This means that we can take 400 information bits and map them into a code of length 1000 bits. Then the whole code can be transmitted over the channels. One hundred of those bits may be detected incorrectly but the 400 information bits may be decoded correctly.

Before we consider continuous-time additive white Gaussian channels, let's concentrate on discrete-time Gaussian channels

Equation:

$$Y_i = X_i + \eta_i$$

where the X_i's are information bearing random variables and η_i is a Gaussian random variable with variance σ_η^2. The input X_i's are constrained to have power less than P

Equation:

$$\frac{1}{n}\sum_{i=1}^{n} {X_i}^2 \leq P$$

Consider an output block of size n

Equation:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\eta}$$

For large n, by the Law of Large Numbers,

Equation:

$$\frac{1}{n}\sum_{i=1}^{n} {\eta_i}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(|y_i - x_i|\right)^2 \leq {\sigma_\eta}^2$$

This indicates that with large probability as n approaches infinity, $\boldsymbol{Y}$ will be located in an n-dimensional sphere of radius $\sqrt{n{\sigma_\eta}^2}$ centered about $\boldsymbol{X}$ since $\left(|\boldsymbol{y} - \boldsymbol{x}|\right)^2 \leq n{\sigma_\eta}^2$

On the other hand since X_i's are power constrained and η_i and X_i's are independent

Equation:

$$\frac{1}{n}\sum_{i=1}^{n} {y_i}^2 \leq P + {\sigma_\eta}^2$$

Equation:

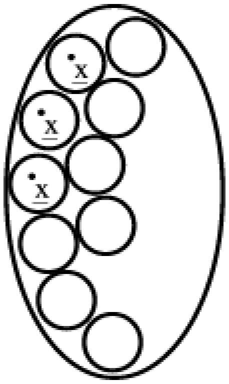
$$|\boldsymbol{Y}| \leq n\left(P+{\sigma_{\eta}}^{2}\right)$$

This mean $\boldsymbol{Y}$ is in a sphere of radius $\sqrt{n\left(P+{\sigma_{\eta}}^{2}\right)}$ centered around the origin.

How many $\boldsymbol{X}$'s can we transmit to have nonoverlapping $\boldsymbol{Y}$ spheres in the output domain? The question is how many spheres of radius $\sqrt{n{\sigma_{\eta}}^{2}}$ fit in a sphere of radius $\sqrt{n\left(P+{\sigma_{\eta}}^{2}\right)}$.

Equation:

$$\begin{aligned} M &= \frac{\left(\sqrt{n\left({\sigma_{\eta}}^{2}+P\right)}\right)^{n}}{\left(\sqrt{n{\sigma_{\eta}}^{2}}\right)^{n}} \\ &= \left(1+\frac{P}{{\sigma_{\eta}}^{2}}\right)^{\frac{n}{2}} \end{aligned}$$



Exercise:

Problem:

How many bits of information can one send in n uses of the channel?

Solution:
Equation:

$$\log_2\left(1+\frac{P}{{\sigma_\eta}^2}\right)^{\frac{n}{2}}$$

The capacity of a discrete-time Gaussian channel $C=\frac{1}{2}\log_2\left(1+\frac{P}{{\sigma_\eta}^2}\right)$ bits per channel use.

When the channel is a continuous-time, bandlimited, additive white Gaussian with noise power spectral density $\frac{N_0}{2}$ and input power constraint P and bandwidth W. The system can be sampled at the Nyquist rate to provide power per sample P and noise power

Equation:

$$\begin{aligned}{\sigma_\eta}^2 &= \int_{-W}^{W}\frac{N_0}{2}\,\mathrm{d}f\\ &= WN_0\end{aligned}$$

The channel capacity $\frac{1}{2}\log_2\left(1+\frac{P}{N_0W}\right)$ bits per transmission. Since the sampling rate is $2W$, then

Equation:

$$C=\frac{2W}{2}\log_2\left(1+\frac{P}{N_0W}\right)\text{bits/trans. x trans./sec}$$

Equation:

$$C=W\log_2\left(1+\frac{P}{N_0W}\right)\frac{\text{bits}}{\text{sec}}$$

Example:

The capacity of the voice band of a telephone channel can be determined using the Gaussian model. The bandwidth is 3000 Hz and the signal to noise ratio is often 30 dB. Therefore,

Equation:

$$C = 3000\log_2(1+1000) \simeq 30000\frac{\text{bits}}{\text{sec}}$$

One should not expect to design modems faster than 30 Kbs using this model of telephone channels. It is also interesting to note that since the signal to noise ratio is large, we are expecting to transmit 10 bits/second/Hertz across telephone channels.

Channel Coding

Channel coding is a viable method to reduce information rate through the channel and increase reliability. This goal is achieved by adding redundancy to the information symbol vector resulting in a longer coded vector of symbols that are distinguishable at the output of the channel. Another brief explanation of channel coding is offered in Channel Coding and the Repetition Code. We consider only two classes of codes, block codes and convolutional codes.

Block codes

The information sequence is divided into blocks of length k. Each block is mapped into channel inputs of length n. The mapping is independent from previous blocks, that is, there is no memory from one block to another.

Example:
$k=2$ and $n=5$
Equation:

$$00 \rightarrow 00000$$

Equation:

$$01 \rightarrow 10100$$

Equation:

$$10 \rightarrow 01111$$

Equation:

$$11 \rightarrow 11011$$

information sequence ⇒ codeword (channel input)

A binary block code is completely defined by 2^k binary sequences of length n called codewords.
Equation:

$$\boldsymbol{C} = \{c_1, c_2, \ldots, c_{2^k}\}$$

Equation:

$$c_i \in \{0, 1\}^n$$

There are three key questions,

1. How can one find "good" codewords?
2. How can one systematically map information sequences into codewords?
3. How can one systematically find the corresponding information sequences from a codeword, i.e., how can we decode?

These can be done if we concentrate on linear codes and utilize finite field algebra.

A block code is linear if $\boldsymbol{c}_i \in \boldsymbol{C}$ and $\boldsymbol{c}_j \in \boldsymbol{C}$ implies $\boldsymbol{c}_i \oplus \boldsymbol{c}_j \in \boldsymbol{C}$ where $\oplus$ is an elementwise modulo 2 addition.

Hamming distance is a useful measure of codeword properties
Equation:

$$d_H(\boldsymbol{c}_i, \boldsymbol{c}_j) = \#\text{ of places that they are different}$$

Denote the codeword for information sequence $e_1 = \begin{matrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{matrix}$ by g_1 and $e_2 = \begin{matrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{matrix}$ by g_2,…, and $e_k = \begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{matrix}$ by g_k. Then any information sequence can be expressed as

Equation:

$$\begin{aligned} \boldsymbol{u} &= \begin{matrix} u_1 \\ \vdots \\ u_k \end{matrix} \\ &= \sum_{i=1}^{k} u_i e_i \end{aligned}$$

and the corresponding codeword could be

Equation:

$$\boldsymbol{c} = \sum_{i=1}^{k} u_i g_i$$

Therefore

Equation:

$$\boldsymbol{c} = \boldsymbol{u}G$$

with $\boldsymbol{c} = \{0,1\}^n$ and $\boldsymbol{u} \in \{0,1\}^k$ where $G = \begin{matrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{matrix}$, a kxn matrix and all operations are modulo 2.

Example:
In [link] with

Equation:

$$00 \rightarrow 00000$$

Equation:

$$01 \rightarrow 10100$$

Equation:

$$10 \rightarrow 01111$$

Equation:

$$11 \rightarrow 11011$$

$g_1 = (01111)^T$ and $g_2 = (10100)^T$ and $G = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$

Additional information about coding efficiency and error are provided in Block Channel Coding.

Examples of good linear codes include Hamming codes, BCH codes, Reed-Solomon codes, and many more. The rate of these codes is defined as $\frac{k}{n}$ and these codes have different error correction and error detection properties.

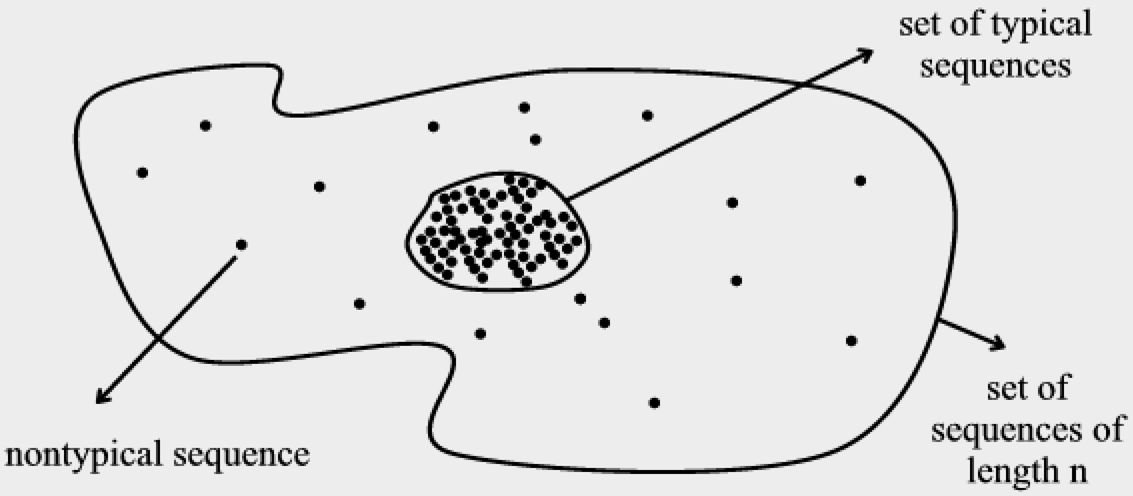
Convolutional Codes

Convolutional codes are one type of code used for channel coding. Another type of code used is block coding.

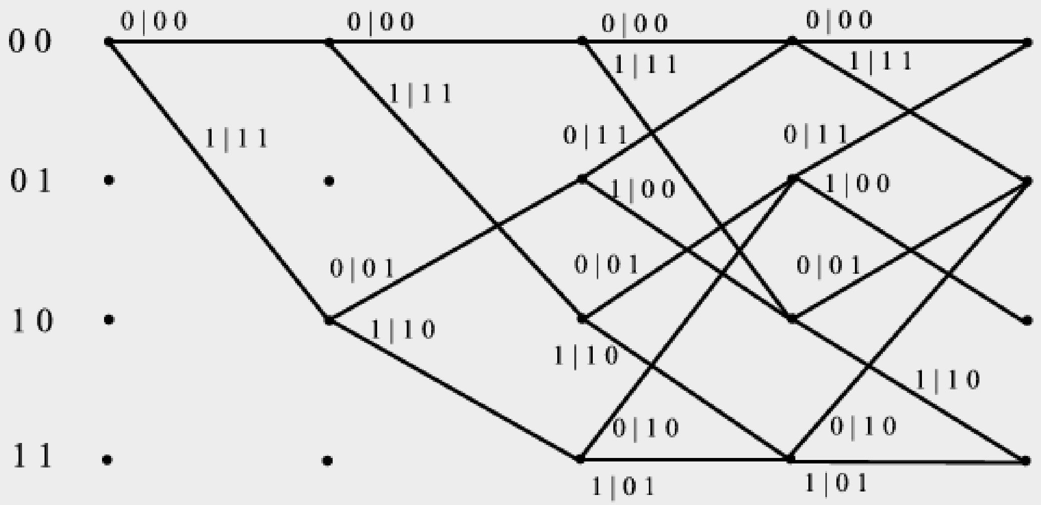
Convolutional codes

In convolutional codes, each block of k bits is mapped into a block of n bits but these n bits are not only determined by the present k information bits but also by the previous information bits. This dependence can be captured by a finite state machine.

Example:
A rate $\frac{1}{2}$ convolutional coder $k = 1, n = 2$ with memory length 2 and constraint length 3.



Since the length of the shift register is 2, there are 4 different rates. The behavior of the convolutional coder can be captured by a 4 state machine.
States: `00, 01, 10, 11`,
For example, arrival of information bit `0` transitions from state `10` to state `01`.
The encoding and the decoding process can be realized in trellis structure.



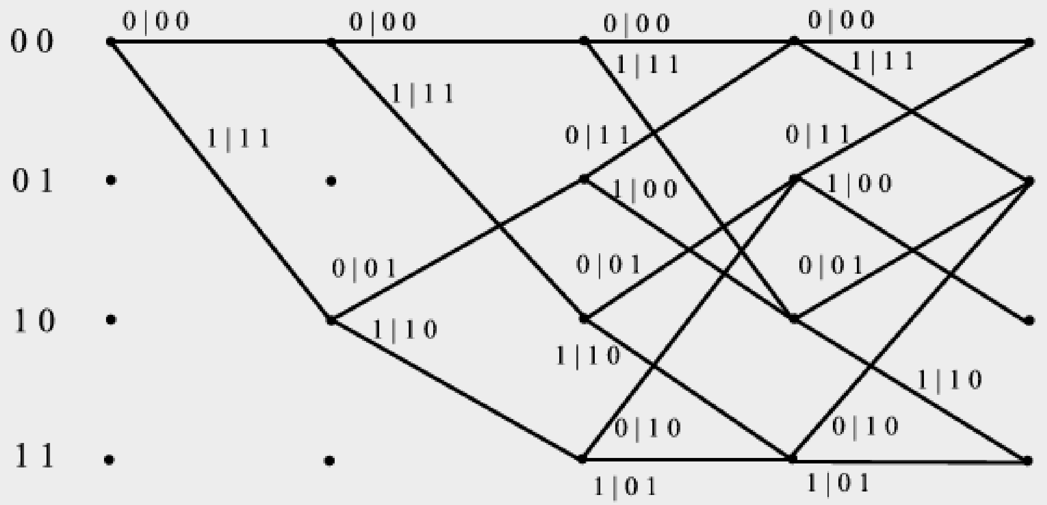
If the input sequence is

```
1 1 0 0
```

the output sequence would be

```
11 10 10 11
```

The transmitted codeword is then `11 10 10 11`. If there is one error on the channel `11 00 10 11`



Starting from state `00` the Hamming distance between the possible paths and the received sequence is measured. At the end, the path with minimum distance to the received sequence is chosen as the correct trellis path. The information sequence will then be determined.

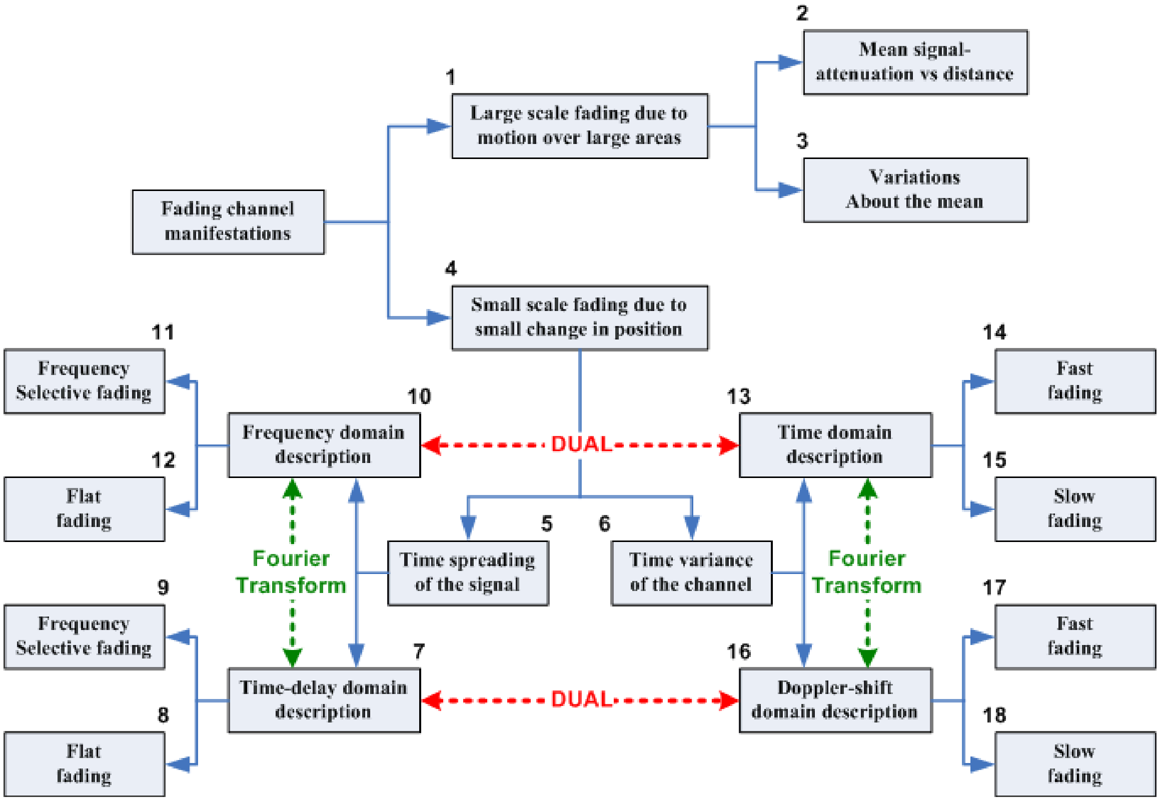
Convolutional coding lends itself to very efficient trellis based encoding and decoding. They are very practical and powerful codes.

Fading Channel

For most channels, where signal propagate in the atmosphere and near the ground, the free-space propagation model is inadequate to describe the channel behavior and predict system performance. In wireless system, s signal can travel from transmitter to receiver over multiple reflective paths. This phenomenon, called multipath fading, can cause fluctuations in the received signal's amplitude, phase, and angle of arrival, giving rise to the terminology multipath fading. Another name, scintillation, is used to describe the fading caused by physical changes in the propagating medium, such as variations in the electron density of the ionosopheric layers that reflect high frequency radio signals. Both fading and scintillation refer to a signal's random fluctuations.

Characterizing Mobile-Radio Propagation

Characterizing Mobile-Radio Propagation

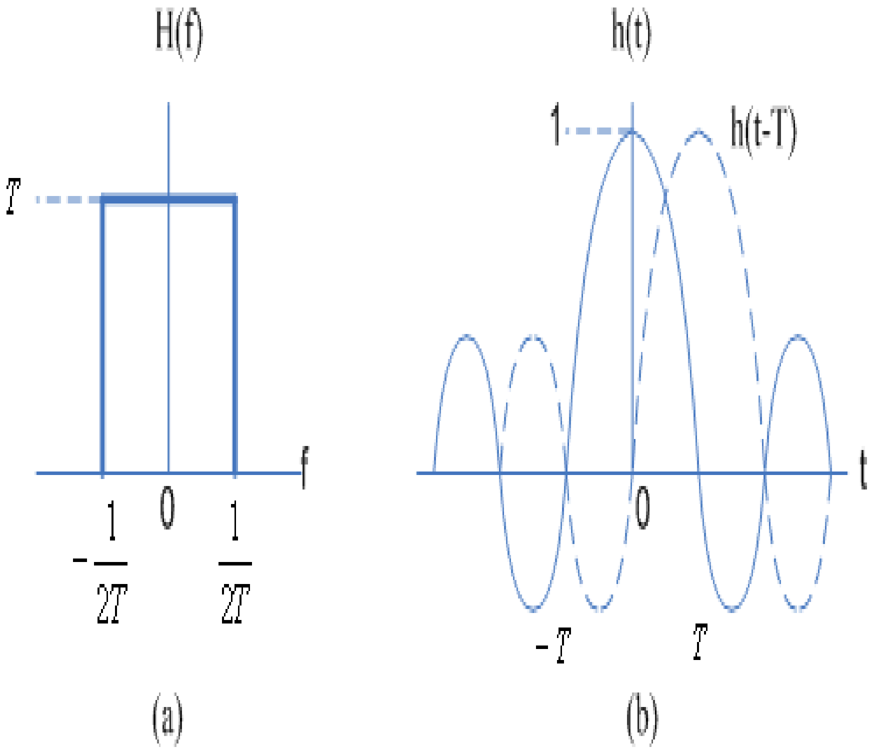

2
Mean signal-attenuation vs distance
1
Large scale fading due to motion over large areas
3
Variations About the mean
Fading channel manifestations
4
Small scale fading due to small change in position
11
Frequency Selective fading
10
Frequency domain description
DUAL
13
Time domain description
14
Fast fading
12
Flat fading
15
Slow fading
5
6
Fourier Transform
Time spreading of the signal
Time variance of the channel
Fourier Transform
9
17
Frequency Selective fading
Fast fading
7
16
Time-delay domain description
DUAL
Doppler-shift domain description
8
18
Flat fading
Slow fading

Fading channel manifestations

Figure 1 introduces an overview of fading channel. Large-scale fading represents the average power attenuation or the path loss due to motion over large areas. This phenomenon is affected by prominent terrain contours (e.g. hills, forests, billboards, clumps of buildings, etc) between the transmitter and receiver. Small-scale fading refers to the dramatic changes in signal amplitude and phase as a result of small changes (as small as half wavelength) in the spatial positioning between a receiver and transmitter. Small-scale fading is called Rayleigh fading if there are multiple reflective paths and no line-of-sight signal component otherwise it is called Rician. When a mobile radio roams over a large area it must process signals that experience both types of fading: small-scale fading superimposed on large-scale fading. Large-scale fading (attenuation or path loss) can be considered as a spatial average over the small-scale fluctuations of the signal.

There are three basic mechanisms that impact signal propagation in a mobile communication system:

1. Reflection occurs when a propagating electromagnetic wave impinges upon smooth surface with very large dimensions relative to the RF signal wavelength.
2. Diffraction occurs when the propagation path between the transmitter and receiver is obstructed by a dense body with dimensions that are large relative to the RF signal wavelength. Diffraction accounts for RF energy traveling from transmitter to receiver without line-of-sight path. It is often termed shadowing because the diffracted field can reach the receiver even when shadowed by an impenetrable obstruction.
3. Scattering occurs when a radio wave impinges on either a large, rough surface or any surface whose dimension are on the other of the RF signal wavelength or less, causing the energy to be spread out or reflected in all directions.



Link budget considerations for a fading channel

Figure 2 is a convenient pictorial showing the various contributions that must be considered when estimating path loss for link budget analysis in a mobile radio application: (1) mean path loss as a function of distance, due to large-scale fading, (2) near-worst-case variations about the mean path loss or large-scale fading margin (typically 6-10 dB), (3) near-worst-case Rayleigh or small-scale fading margin (typically 20-30 dB)

Using complex notation

$s(t) = \mathrm{Re}\left\{g(t).e^{j2\pi f_c t}\right\}$(1)

Where $\mathrm{Re}\{.\}$ denotes the real part of $\{.\}$, and f_c is the carrier frequency. The baseband waveform $g(t)$ is called the complex envelope of $s(t)$ and can be expressed as

$g(t) = \mid g(t) \mid .e^{j\varphi(t)} = R(t).e^{j\varphi(t)}$(2)

Where $R(t) = \mid g(t) \mid$ is the envelope magnitude, and $\varphi(t)$ is its phase.

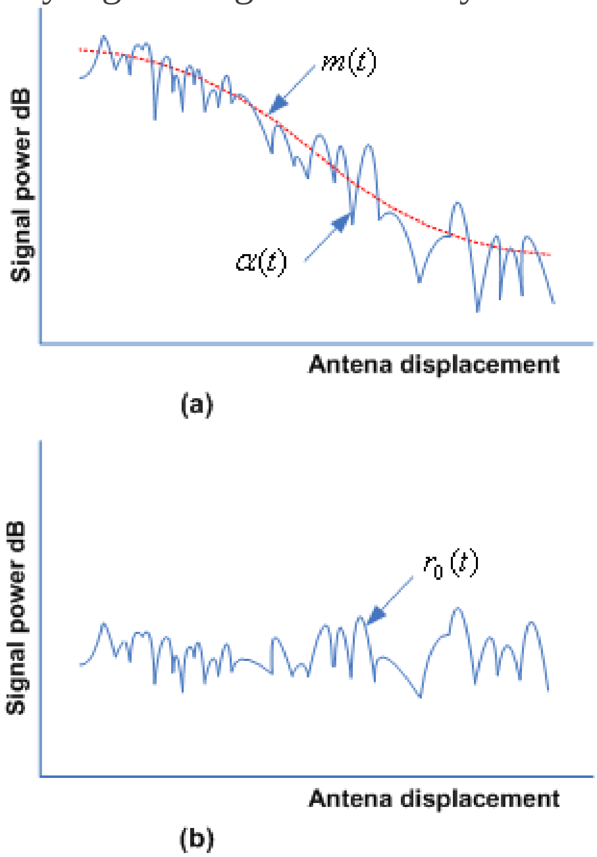
In fading environment, g(t) will be modified by a complex dimentionless multiplicative factor $\alpha(t).e^{-j\theta(t)}$. The modified baseband waveform can be written as $\alpha(t).e^{-j\theta(t)}.g(t)$. The magnitude of this envelope can be expressed as follow

$\alpha(t).R(t) = m(t).r_0(t).R(t)$(3)

Where $m(t)$ and $r_0(t)$ are called the large-scale-fading component and the large-scale-fading component of the envelope respectively.

Sometimes, $m(t)$ is referred to as the local mean or log-normal fading, and $r_0(t)$ is referred to as multipath or Rayleigh fading.

For the case of mobile radio, figure 3 illustrates the relationship between $\alpha(t).m(t)$. In figure 3a, the signal power received is a function of the multiplicative factor $\alpha(t)$. Small-scale fading superimposed on large-scale fading can be readily identified. The typical antenna displacement between adjacent signal-strength nulls due to small-scale fading is approximately half of wavelength. In figure 3b, the large-scale fading or local mean $m(t)$ has been removed in order to view the small-scale fading $r_0(t)$. The log-normal fading is a relative slow varying function of position, while the Rayleigh fading is a relatively fast varying function of position.



Large-scale fading and small-scale fading

Large-Scale Fading

In general, propagation models for both indoor and outdoor radio channels indicate that mean path loss as follow

$L_p(d)$~ $d/d_0\ ^n$(1)

$L_p(d)\mathrm{dB} = L_s(d_0)\mathrm{dB} + 10n.\log(d/d_0)$ (2)

Where d is the distance between transmitter and receiver, and the reference distance d_0 corresponds to a point located in the far field of the transmit antenna. Typically, d_0 is taken 1 km for large cells, 100 m for micro cells, and 1 m for indoor channels. Moreover d_0 is evaluated using equation

$L_s(d_0) = \ \frac{4\pi \mathrm{d}_0}{\lambda}\ ^2$(3)

or by conducting measurement. The value of the path-loss exponent n depends on the frequency, antenna height and propagation environment. In free space, n is equal to 2. In the presence of a very strong guided wave phenomenon (like urban streets), n can be lower than 2. When obstructions are present, n is larger.

Measurements have shown that the path loss L_p is a random variable having a log-normal distribution about the mean distant-dependent value $L_p(d)$

$L_p(d)(\mathrm{dB}) = L_s(d_0)(\mathrm{dB}) + 10n\log_{10}(d/d_0) + X_\sigma(\mathrm{dB})$(4)

Where X_σ denote a zero-mean, Gaussian random variable (in dB) with standard deviation   (in dB). X_σ is site and distance dependent.

As can be seen from the equation, the parameters needed to statistically describe path loss due to large-scale fading, for an arbitrary location with a specific transmitter-receiver separation are (1) the reference distance, (2) the path-loss exponent, and (3) the standard deviation X_σ.

Small-Scale Fading

SMALL - SCALE FADING

Small-scale fading refers to the dramatic changes in signal amplitude and phase that can be experienced as a result of small changes (as small as half wavelength) in the spatial position between transmitter and receiver.

In this section, we will develop the small-scale fading component $r_0(t)$. Analysis proceeds on the assumption that the antenna remains within a limited trajectory so that the effect of large-scale fading m(t) is constant. Assume that the antenna is traveling and there are multiple scatter paths, each associated with a time-variant propagation delay $\tau_n(t)$ and a time variant multiplicative factor $\alpha_n(t)$. Neglecting noise, the received bandpass signal can be written as below:

$$r(t) = \sum_n \alpha_n(t) s(t - \tau_n(t))(1)$$

Substituting Equation (1, module Characterizing Mobile-Radio Propagation) over into Equation (1), we can write the received bandpass signal as follow:

$$r(t)\text{=Re}((\sum_n \alpha_n(t) g(t - \tau_n(t)) e^{j2\pi f_c(t-\tau_n(t))})(2)$$

$$= \text{Re}((\sum_n \alpha_n(t) e^{-j2\pi f_c \tau_n(t)} g(t - \tau_n(t))) e^{j2\pi f_c t}$$

We have the equivalent received bandpass signal is

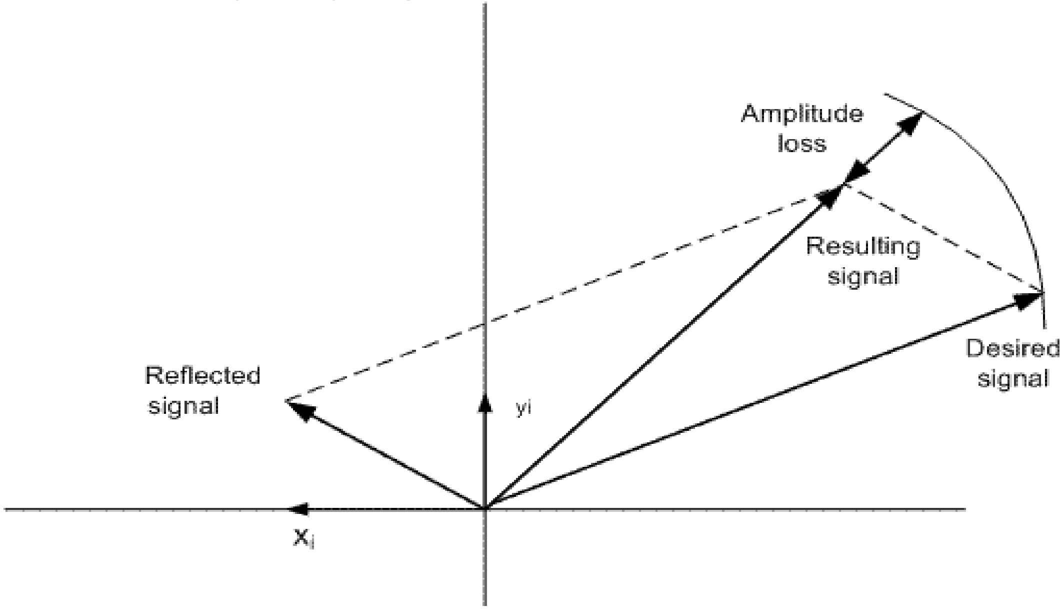
$$s(t) = \sum_n \alpha_n(t) e^{-j2\pi f \tau_n(t)_c} g(t - \tau_n(t))(3)$$

Consider the transmission of an unmodulated carrier at frequency f_c or in other words, for all time, g(t)=1. So the received bandpass signal become as follow:

$$s(t) = \sum_n \alpha_n(t) e^{-j2\pi f_c \tau_n(t)} = \sum_n \alpha_n(t) e^{-j\theta_n(t)}(4)$$

The baseband signal s(t) consists of a sum of time-variant components having amplitudes $\alpha_n(t)$ and phases $\theta_n(t)$. Notice that $\theta_n(t)$ will change by

2π radians whenever $\tau_n(t)$ changes by 1/ f_c (very small delay). These multipath components combine either constructively or destructively, resulting in amplitude variations of s(t). Final equation is very important because it tell us that a bandpass signal s(t) is the signal that experienced the fading effects and gave rise to the received signal r(t), these effects can be described by analyzing r(t) at the baseband level.



When the received signal is made up of multiple reflective arrays plus a significant line-of-sight (non-faded) component, the received envelope amplitude has a Rician pdf as below, and the fading is preferred to as Rician fading

$$p(r_0) = \begin{cases} \frac{r_0}{\sigma^2}\exp\left[-\frac{(r_0^2+A^2)}{2\sigma^2}\right] I_0\left(\frac{r_0 A}{\sigma^2}\right) & r_0 \geq 0, A \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

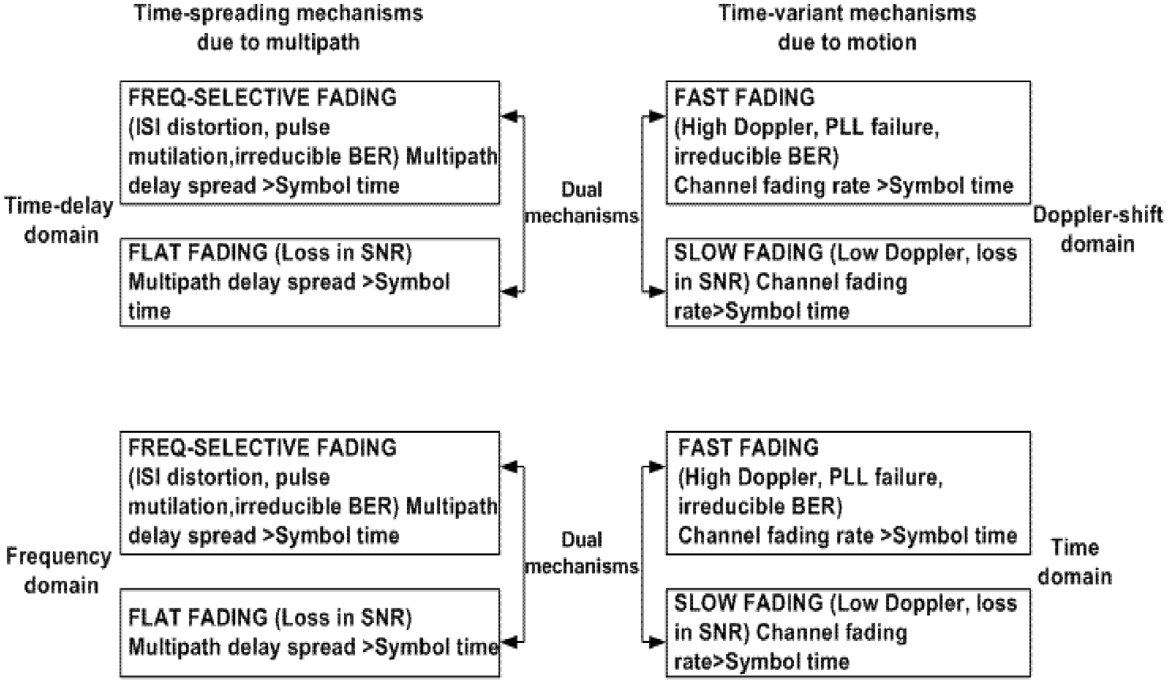
The parameter σ^2 is the pre-detection mean power of the multipath signal. A denotes the peak magnitude of the non-faded signal component and $I_0(-)$ is the modified Bessel function. The Rician distribution is often described in terms of a parameter K, which is defined as the ratio of the power in the specular component to the power in the multipath signal. It is given by $K = A^2/2\sigma^2$.

When the magnitude of the specular component A approach zero, the Rician pdf approachs a Rayleigh pdf, shown as

$$p(r_0) = \begin{cases} \frac{r_0}{\sigma^2} \exp\left[-\frac{r_0^2}{2\sigma^2}\right] & r_0 \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The Rayleigh pdf results from having no specular signal component, it represents the pdf associated with the worst case of fading per mean received signal power.

Small scale manifests itself in two mechanisms - time spreading of signal (or signal dispersion) and time-variant behavior of the channel (figure 2). It is important to distinguish between two different time references- delay time τ and transmission time t. Delay time refers to the time spreading effect resulting from the fading channel’s non-optimum impulse response. The transmission time, however, is related to the motion of antenna or spatial changes, accounting for propagation path changes that are perceived as the channel’s time-variant behavior.


Time-spreading mechanisms due to multipath
Time-variant mechanisms due to motion
FREQ-SELECTIVE FADING (ISI distortion, pulse mutilation,irreducible BER) Multipath delay spread >Symbol time
FAST FADING (High Doppler, PLL failure, irreducible BER) Channel fading rate >Symbol time
Time-delay domain
Dual mechanisms
Doppler-shift domain
FLAT FADING (Loss in SNR) Multipath delay spread >Symbol time
SLOW FADING (Low Doppler, loss in SNR) Channel fading rate>Symbol time
FREQ-SELECTIVE FADING (ISI distortion, pulse mutilation,irreducible BER) Multipath delay spread >Symbol time
FAST FADING (High Doppler, PLL failure, irreducible BER) Channel fading rate >Symbol time
Frequency domain
Dual mechanisms
Time domain
FLAT FADING (Loss in SNR) Multipath delay spread >Symbol time
SLOW FADING (Low Doppler, loss in SNR) Channel fading rate>Symbol time

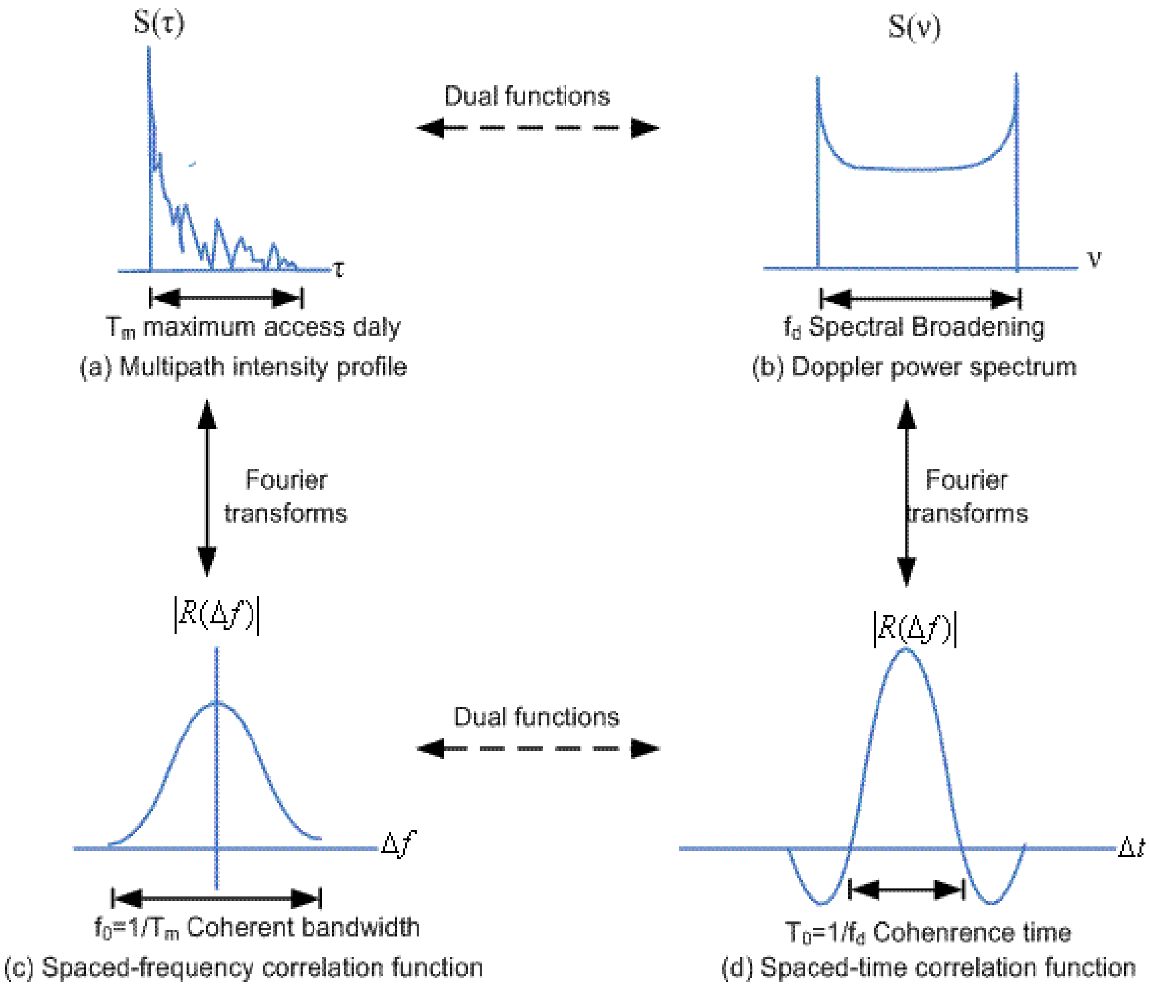
Signal Time-Spreading

SIGNAL TIME – SPREADING

Signal Time-Spreading Viewed in the Time-Delay Domain

A simple way to model the fading phenomenon is proposed the notion wide-sense stationary uncorrelated scattering. The model treats arriving at a receive antenna with different delay as uncorrelated.

In Figure 1(a), a multipath-intensity profile S(τ) is plotted. S(τ) helps us understand how the average received power vary as a function of time delay τ. The term “time delay” is used to refer to the excess delay. It represents the signal’s propagation delay that exceeds the delay of the first signal arrival at the receiver. In wireless channel, the received signal usually consists of several discrete multipath components causing S(τ). For a single transmitted impulse, the time T_m between the first and last received component represents the maximum excess delay.


S(τ)
Dual functions
S(ν)
τ
ν
Tm maximum access daly
(a) Multipath intensity profile
fd Spectral Broadening
(b) Doppler power spectrum
Fourier transforms
Fourier transforms
|R(Δf)|
|R(Δf)|
Dual functions
Δf
Δt
f0=1/Tm Coherent bandwidth
(c) Spaced-frequency correlation function
T0=1/fd Cohenrence time
(d) Spaced-time correlation function

Degradation Categories due to Signal Time-Spreading Viewed in the Time-Delay Domain

In a fading channel, the relationship between maximum excess delay time T_m and symbol time T_s can be viewed in terms of two different degradation categories: frequency-selective fading and frequency nonselective or flat fading.

A channel is said to exhibit frequency selective fading if $T_m > T_s$. This condition occurs whenever the received multipath components of a symbol extend beyond the symbol's time duration. In fact, another name for this category of fading degradation is channel-induced ISI. In this case of frequency-selective fading, mitigating the distortion is possible because many of the multipath components are resolved by receiver.

A channel is said to exhibit frequency nonselective or flat fading if $T_m < T_s$. In this case, all of the received multipath components of a symbol arrive within the symbol time duration; hence, the components are not resolvable. There is no channel-induced ISI distortion because the signal time spreading does not result in significant overlap among neighboring received symbols.

Signal Time-Spreading Viewed in the Frequency Domain

A completely analogous characterization of signal dispersion can be specified in the frequency domain. In figure 1b, the spaced-frequency correlation function $\mid R(\mathit{\Delta} f) \mid$ can be seen, it is the Fourier transform of S(τ). The correlation function $\mid R(\mathit{\Delta} f) \mid$ represents the correlation between the response of channel to two signals as a function of the frequency difference between two signals. The function $\mid R(\mathit{\Delta} f) \mid$ helps answer the correlation between received signals that are spaced in the frequency $\mathit{\Delta} f = f_1 - f_2$ is what. $\mid R(\mathit{\Delta} f) \mid$ can be measured by transmitting a pair of sinusoids separated in frequency by Δf, cross-correlating the complex spectra of two separated received signals, and repeating the process many times with ever-larger separation Δf. Spectral components in that range are affected by the channel in a similar manner. Note that the coherence bandwidth f_0 and the maximum excess delay time T_m are related as approximation below

$f_0 \approx \frac{1}{T_m}$(1)

A more useful parameter is the delay spread, most often characterized in terms of its root-mean-square (rms) value, can be calculated as

$\sigma_\tau = (\tau^2 - \tau^2)^{1/2}$(2)

Where $\bar{\tau}$ is the mean excess delay, $(\bar{\tau})^2$ is the mean squared, τ^2 is the second moment and σ_τ is the square root of the second central moment of S(τ).

A relationship between coherence bandwidth and delay spread does not exist. However, using Fourier transform techniques an approximation can be derived from actual signal dispersion measurements in various channel. Several approximate relationships have been developed.

If the coherence bandwidth is defined as the frequency interval over which the channel’s complex frequency transfer function has a correlation of at least 0.9, the coherent bandwidth is approximately

$f_0 \approx \frac{1}{50\sigma_\tau}$(3)

With the dense-scatterer channel model, coherence bandwidth is defined as the frequency interval over which the channel’s complex frequency transfer function has a correlation of at least 0.5, to be

$f_0 \approx \frac{1}{2\pi\sigma_\tau}$(4)

Studies involving ionospheric effects often employ the following definition

$f_0 \approx \frac{1}{5\sigma_\tau}$(5)

The delay spread and coherence bandwidth are related to a channel’s multipath characteristic, differing for different propagation paths. It is important to note that all parameters in last equation independent of signaling speed, a system’s signaling speed only influences its transmission bandwidth W.

Degradation Categories due to Signal Time-Spreading Viewed in the Frequency Domain

A channel is preferred to as frequency-selective if $f_0 < 1/T_s \approx W$ (the symbol rate is taken to be equal to the signaling rate or signal bandwidth W). Frequency selective fading distortion occurs whenever a signal’s spectral components are not all affected equally by the channel. Some of the signal’s spectra components failing outside the coherent bandwidth will be affected differently, compared with those components contained within the coherent bandwidth (Figure 2(a)).

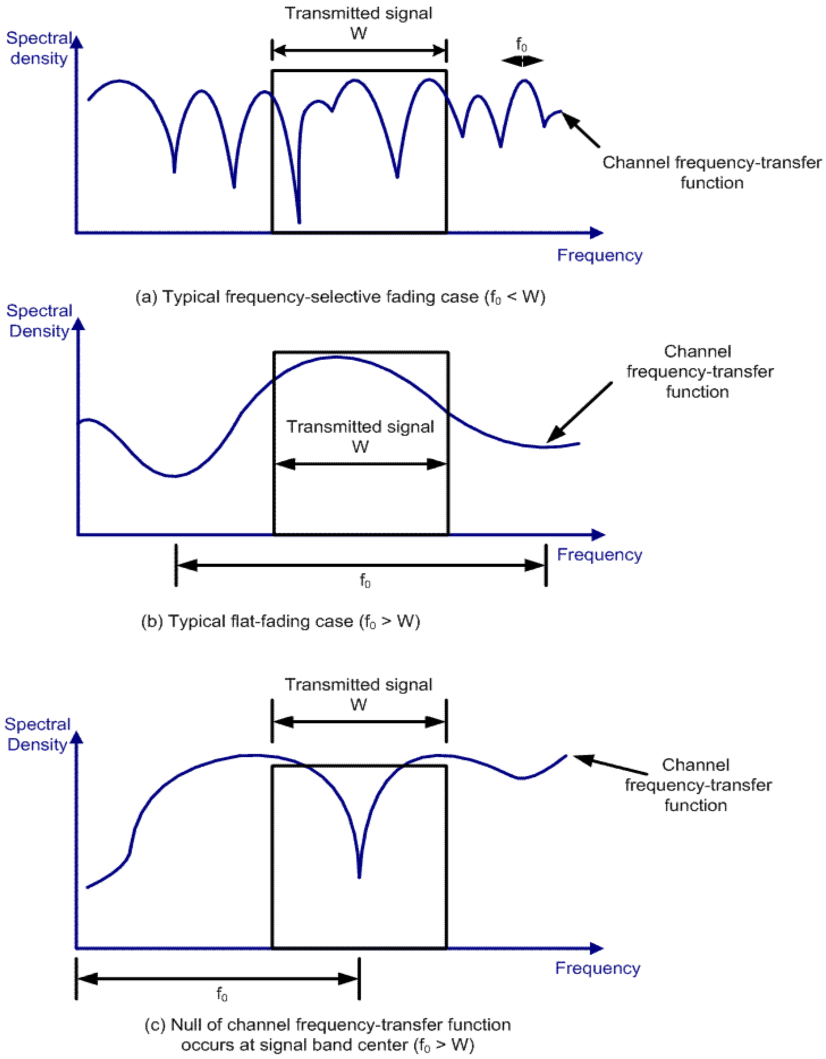
Frequency- nonselective of flat-fading degradation occurs whenever $f_0 > W$. hence, all of signal’s spectral components will be affected by the channel in a similar manner (fading or non-fading) (Figure 2(b)). Flat fading does not introduce channel-induced ISI distortion, but performance degradation can still be expected due to the loss in SNR whenever the signal is fading. In order to avoid channel-induced ISI distortion, the channel is required to exhibit flat fading. This occurs, provide that

$$f_0 > W \approx \frac{1}{T_s}$$

(6)

Hence, the channel coherent bandwidth f0 set an upper limit on the transmission rate that can be used without incorporating an equalizer in the receiver.

However, as a mobile radio changes its position, there will be times when the received signal experiences frequency-selective distortion even though $f_0 > W$ (in Figure 2(c)). When this occurs, the baseband pulse can be especially mutilated by deprivation of its low-frequency components. Thus, even though a channel is categorized as flat-fading, it still manifests frequency-selective fading.



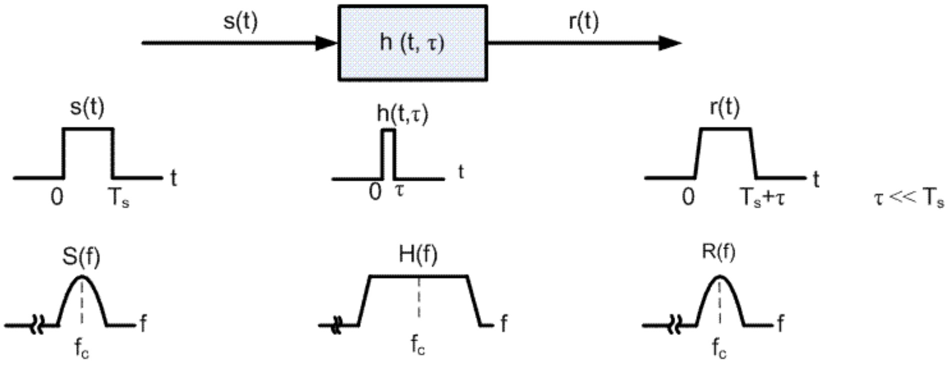
(a) Typical frequency-selective fading case ($f_0 < W$)

(b) Typical flat-fading case ($f_0 > W$)

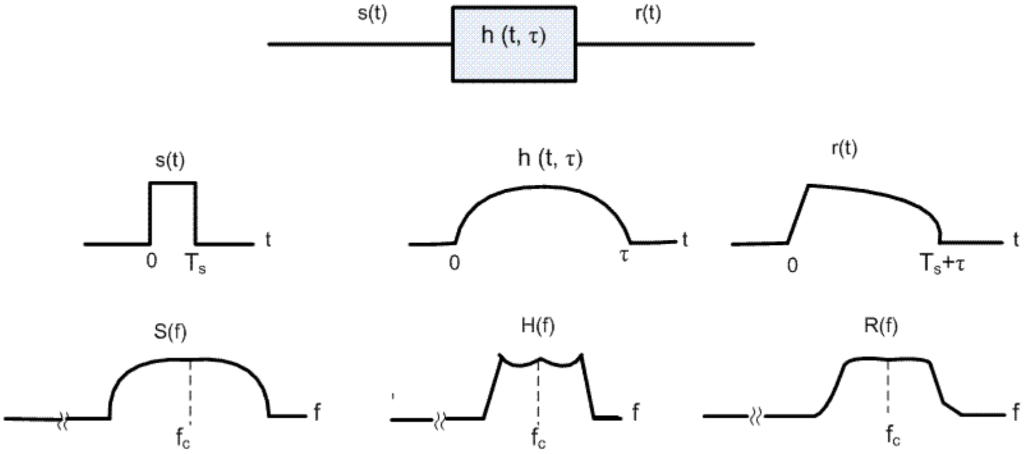
(c) Null of channel frequency-transfer function occurs at signal band center ($f_0 > W$)

Examples of Flat Fading and Frequency-Selective Fading

The signal dispersion manifestation of the fading channel is analogous to the signal spreading that characterizes an electronic filter. Figure 3(a) depicts a wideband filter (narrow impulse response) and its effect on a signal in both time domain and the frequency domain. This filter resembles a flat-fading channel yielding an output that is relatively free of dispersion. Figure 3(b) shows a narrowband filter (wide impulse response). The output signal suffers much distortion, as shown both time domain and frequency domain. Here the process resembles a frequency-selective channel.



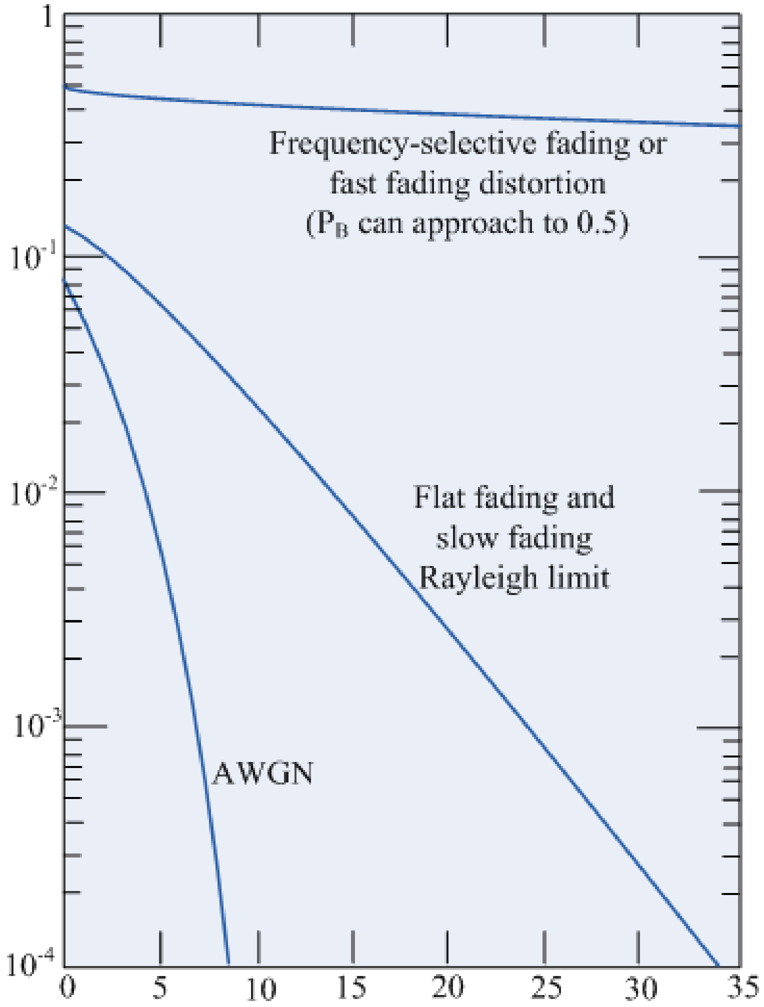
(a) Flat fading channel characteristics



(b) Frequency selective fading channel characteristics

Mitigating the Degradation Effects of Fading

Figure 1 highlights three major performance categories in terms of bit-error probability $P_B$versus E_b N



The leftmost exponentially shaped curve highlights the performance that can be expected when using any nominal modulation scheme in AWGN interference. Observe that at a reasonable E_b N level, good performance can be expected.

The middle curve, referred to as the **Rayleigh limit**, shows the performance degradation resulting from a loss in E_b N that is characteristic of flat fading or slow fading when there is no line-of-sight signal component present. The curve is a function of the reciprocal of E_b N , so for practical values of E_b N , performance will generally be “bad.”

The curve that reaches an irreducible error-rate level, sometimes called an **error floor**, represents “awful” performance, where the bit-error probability can level off at values nearly equal to 0.5. This shows the severe performance degrading effects that are possible with frequency-selective fading or fast fading.

If the channel introduces signal distortion as a result of fading, the system performance can exhibit an irreducible error rate at a level higher than the desired error rate. In such cases, the only approach available for improving performance is to use some forms of mitigation to remove or reduce the signal distortion.

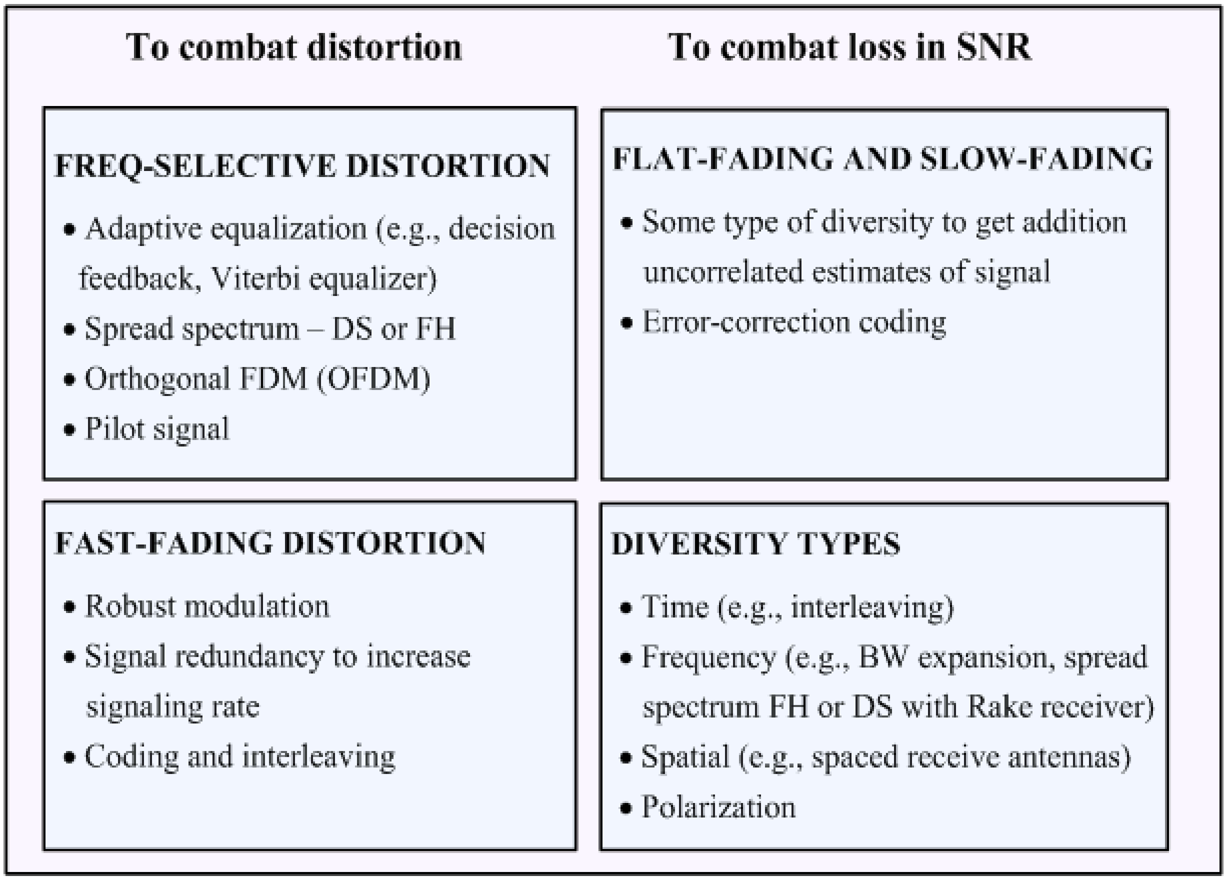
The mitigation method depends on whether the distortion is caused by frequency-selective fading or fast fading. Once the signal distortion has been mitigated, the P_B versus $E_b \;\; N$ performance can transition from the “awful” category to the merely “bad” Rayleigh-limit curve.

Next, it is possible to further ameliorate the effects of fading and strive to approach AWGN system performance by using some form of diversity to provide the receiver with a collection of uncorrelated replicas of the signal, and by using a powerful error-correction code.

Figure 2 lists several mitigation techniques for combating the effects of both signal distortion and loss in SNR. The mitigation approaches to be used when designing a system should be considered in two basic steps:

1) choose the type of mitigation to reduce or remove any distortion degradation;

2) choose a diversity type that can best approach AWGN system performance.



To combat distortion

To combat loss in SNR

FREQ-SELECTIVE DISTORTION

- Adaptive equalization (e.g., decision feedback, Viterbi equalizer)
- Spread spectrum – DS or FH
- Orthogonal FDM (OFDM)
- Pilot signal

FLAT-FADING AND SLOW-FADING

- Some type of diversity to get addition uncorrelated estimates of signal
- Error-correction coding

FAST-FADING DISTORTION

- Robust modulation
- Signal redundancy to increase signaling rate
- Coding and interleaving

DIVERSITY TYPES

- Time (e.g., interleaving)
- Frequency (e.g., BW expansion, spread spectrum FH or DS with Rake receiver)
- Spatial (e.g., spaced receive antennas)
- Polarization

Mitigation to Combat Frequency-Selective Distortion

Equalization can mitigate the effects of channel-induced ISI brought on by frequency-selective fading. It can help modify system performance described by the curve that is “awful” to the one that is merely “bad.” The process of equalizing for mitigating ISI effects involves using methods to gather the dispersed symbol energy back into its original time interval.

An equalizer is an inverse filter of the channel. If the channel is frequency selective, the equalizer enhances the frequency components with small amplitudes and attenuates those with large amplitudes. The goal is for the combination of channel and equalizer filter to provide a flat composite-received frequency response and linear phase.

Because the channel response varies with time, the equalizer filters must be **adaptive equalizers**.

The **decision feedback equalizer (DFE)** involves:

1) a feedforward section that is a linear transversal filter whose stage length and tap weights are selected to coherently combine virtually all of the current symbol’s energy.

2) a feedback section that removes energy remaining from previously detected symbols.

The basic idea behind the DFE is that once an information symbol has been detected, the ISI that it induces on future symbols can be estimated and subtracted before the detection of subsequent symbols.

A **maximum-likelihood sequence estimation (MLSE)** equalizer: tests all possible data sequences and chooses the data sequence that is the most probable of all the candidates. The MLSE is optimal in the sense that it minimizes the probability of a sequence error. Since the MLSE equalizer is implemented by using **Viterbi decoding algorithm**, it is often referred to as the **Viterbi equalizer**.

Direct-sequence spread-spectrum (DS/SS) techniques can be used to mitigate frequency-selective ISI distortion because the hallmark of spread-spectrum systems is their capability of rejecting interference, and ISI is a type of interference.

Consider a DS/SS binary **phase-shift keying (PSK)** communication channel comprising one direct path and one reflected path. Assume that the propagation from transmitter to receiver results in a multipath wave that is delayed by τ compared to the direct wave. The received signal, $r(t)$, neglecting noise, can be expressed as follows:

$$r(t) = \mathrm{Ax}(t)g(t)\cos(2\pi \mathrm{f}_c t) + \alpha \mathrm{Ax}(t-\tau)g(t-\tau)\cos(2\pi \mathrm{f}_c t + \theta)$$

where $x(t)$ is the data signal, $g(t)$ is the **pseudonoise (PN)** spreading code, and τ is the differential time delay between the two paths. The angle θ is a random phase, assumed to be uniformly distributed in the range $(0,2\pi)$, and α is the attenuation of the multipath signal relative to the direct path signal.

The receiver multiplies the incoming $r(t)$ by the code $g(t)$. If the receiver is synchronized to the direct path signal, multiplication by the code signal yields the following:

$$r(t)g(t) = \mathrm{Ax}(t)g^2(t)\cos(2\pi \mathrm{f}_c t) + \alpha \mathrm{Ax}(t-\tau)g(t)g(t-\tau)\cos(2\pi \mathrm{f}_c t + \theta)$$

where $g^2(t) = 1$. If τ is greater than the chip duration, then

$$\mid \int g(t)g(t-\tau)\mathrm{dt} \mid \leq \mid \int g^2(t)\mathrm{dt} \mid$$

over some appropriate interval of integration (correlation). Thus, the spread spectrum system effectively eliminates the multipath interference by virtue of its code-correlation receiver. Even though channel-induced ISI is typically transparent to DS/SS systems, such systems suffer from the loss in energy contained in the multipath components rejected by the receiver. The need to gather this lost energy belonging to a received chip was the motivation for developing the **Rake receiver**.

A channel that is classified as flat fading can occasionally exhibit frequency-selective distortion when the null of the channel’s frequency-transfer function occurs at the center of the signal band. The use of DS/SS is a practical way of mitigating such distortion because the wideband SS signal can span many lobes of the selectively faded channel frequency response. This requires the spread-spectrum bandwidth W_{ss} (or the chip rate R_{ch}), to be greater than the coherence bandwidth f_0. The larger the ratio of W_{ss} to f_0 , the more effective will be the mitigation.

Frequency-hopping spread-spectrum (FH/SS): can be used to mitigate the distortion caused by frequency-selective fading, provided that the hopping rate is at least equal to the symbol rate. FH receivers avoid the degradation effects due to multipath by rapidly changing in the transmitter carrier-frequency band, thus avoiding the interference by changing the receiver band position before the arrival of the multipath signal.

Orthogonal frequency-division multiplexing (OFDM): can be used for signal transmission in frequency-selective fading channels to avoid the use of an equalizer by lengthening the symbol duration. The approach is to partition (demultiplex) a high symbol-rate sequence into N symbol groups, so that each group contains a sequence of a lower symbol rate (by the factor $1/N$) than the original sequence. The signal band is made up of N orthogonal carrier waves, and each one is modulated by a different symbol group. The goal is to reduce the symbol rate (signaling rate), $W \approx 1/T_s$, on each carrier to be less than the channel's coherence bandwidth f_0.

Pilot signal is the name given to a signal intended to facilitate the coherent detection of waveforms. Pilot signals can be implemented in the frequency domain as in-band tones, or in the time domain as digital sequences that can also provide information about the channel state and thus improve performance in fading conditions.

Mitigation to Combat Fast-Fading Distortion

- For fast-fading distortion, use a robust modulation (non-coherent or differentially coherent) that does not require phase tracking, and reduces the detector integration time.
- Increase the symbol rate, $W \approx 1/T_s$, to be greater than the fading rate, $f_d \approx 1/T_0$, by adding signal redundancy.
- Error-correction coding and interleaving can provide mitigation, because instead of providing more signal energy, a code reduces the required E_b/N_0. For a given E_b/N_0 with coding present, the error floor will be lowered compared to the uncoded case.

When fast-fading distortion and frequency-selective distortion occur simultaneously, the frequency-selective distortion can be mitigated by the use of an OFDM signal set. Fast fading, however, will typically degrade conventional OFDM because the Doppler spreading corrupts the orthogonality of the OFDM subcarriers. A polyphase filtering technique is used to provide time-domain shaping and partial-response coding to reduce the spectral sidelobes of the signal set, and thus help preserve its orthogonality. The process introduces known ISI and adjacent channel interference (ACI) which are then removed by a post-processing equalizer and canceling filter.

Mitigation to Combat Loss in SNR

Until this point, we have considered the mitigation to combat frequency-selective and fast-fading distortions. The next step is to use diversity methods to move the system operating point from the error-performance curve labeled as “bad” to a curve that approaches AWGN performance. The term diversity is used to denote the various methods available for providing the receiver with uncorrelated renditions of the signal of interest. Some of the ways in which diversity methods can be implemented are:

- **Time diversity**: transmit the signal on L different time slots with time separation of at least T_0. When used along with error-correction coding, interleaving is a form of time diversity.

- **Frequency diversity**: transmit the signal on L different carriers with frequency separation of at least f_0. Bandwidth expansion is a form of frequency diversity. The signal bandwidth W is expanded so as to be greater than f_0, thus providing the receiver with several independently-fading signal replicas. This achieves frequency diversity of the order $L = W/f_0$.

Whenever W is made larger than f_0, there is the potential for frequency-selective distortion unless mitigation in the form of equalization is provided.

Thus, an expanded bandwidth can improve system performance (via diversity) only if the frequency-selective distortion that the diversity may have introduced is mitigated.

- **Spread spectrum**: In spread-spectrum systems, the delayed signals do not contribute to the fading, but to interchip interference. Spread spectrum is a bandwidth-expansion technique that excels at rejecting interfering signals. In the case of **Direct-Sequence Spread-Spectrum (DS/SS)**, multipath components are rejected if they are time-delayed by more than the duration of one chip. However, in order to approach AWGN performance, it is necessary to compensate for the loss in energy contained in those rejected components. The **Rake receiver** makes it possible to coherently combine the energy from several of the multipath components arriving along different paths (with sufficient differential delay).

• **Frequency-hopping spread-spectrum (FH/SS)** is sometimes used as a diversity mechanism. The GSM system uses slow FH (217 hops/s) to compensate for cases in which the mobile unit is moving very slowly (or not at all) and experiences deep fading due to a spectral null.

• **Spatial diversity** is usually accomplished through the use of multiple receive antennas, separated by a distance of at least 10 wavelengths when located at a base station (and less when located at a mobile unit). Signal-processing techniques must be employed to choose the best antenna output or to coherently combine all the outputs. Systems have also been implemented with multiple transmitters, each at a different location.

• **Polarization diversity** is yet another way to achieve additional uncorrelated samples of the signal.

• Some techniques for improving the loss in SNR in a fading channel are more efficient and more powerful than repetition coding.

Error-correction coding represents a unique mitigation technique, because instead of providing more signal energy it reduces the required E_b/N_0 needed to achieve a desired performance level. Error-correction coding coupled with interleaving is probably the most prevalent of the mitigation schemes used to provide improved system performance in a fading environment.

Diversity Techniques

This section shows the error-performance improvements that can be obtained with the use of diversity techniques.

The bit-error-probability, P_B, averaged through all the “ups and downs” of the fading experience in a slow-fading channel is as follows:

$$P_B = \int P_B(x)p(x)\mathrm{dx}$$

where $P_B(x)$ is the bit-error probability for a given modulation scheme at a specific value of $\mathrm{SNR} = x$, where $x = \alpha^2 E_b/N_0$, and $p(x)$ is the pdf of x due to the fading conditions. With E_b and N_0 constant, α is used to represent the amplitude variations due to fading.

For **Rayleigh fading**, α has a **Rayleigh distribution** so that α^2, and consequently x, have a **chi-squared distribution**:

$$p(x) = \frac{1}{\mathit{\Gamma}}\exp(-\frac{x}{\mathit{\Gamma}})\ x \geq 0$$

where $\mathit{\Gamma} = \alpha^2 E_b/N_0$ is the SNR averaged through the “ups and downs” of fading. If each diversity (signal) branch, $i = 1{,}2{,}\ldots{,}M$, has an instantaneous $\mathrm{SNR} = \gamma_i$, and we assume that each branch has the same average SNR given by $\mathit{\Gamma}$, then

$$p(\gamma_i) = \frac{1}{\mathit{\Gamma}}\exp(-\frac{\gamma_i}{\mathit{\Gamma}})\ \gamma_i \geq 0$$

The probability that a single branch has SNR less than some threshold γ is:

$$P(\gamma_i \leq \gamma) = \int_0^\gamma p(\gamma_i)d\gamma_i = \int_0^\gamma \frac{1}{\mathit{\Gamma}}\exp(-\frac{\gamma_i}{\mathit{\Gamma}})d\gamma_i$$

$$= 1 - \exp(-\frac{\gamma}{\mathit{\Gamma}})$$

The probability that all M independent signal diversity branches are received simultaneously with an SNR less than some threshold value γ is:

$$P(\gamma_1{,}\ldots{,}\gamma_M \leq \gamma) = \left[1 - \exp(-\frac{\gamma}{\mathit{\Gamma}})\right]^M$$

The probability that any single branch achieves $\mathrm{SNR} > \gamma$ is:

$$P(\gamma_i > \gamma) = 1 - \left[1 - \exp(-\tfrac{\gamma}{\mathit{\Gamma}})\right]^M$$

This is the probability of exceeding a threshold when selection diversity is used.

Example: Benefits of Diversity

Assume that four-branch diversity is used, and that each branch receives an independently Rayleigh-fading signal. If the average SNR is $\mathit{\Gamma} = 20\mathrm{dB}$, determine the probability that all four branches are received simultaneously with an SNR less than $10\mathrm{dB}$ (and also, the probability that this threshold will be exceeded).

Compare the results to the case when no diversity is used.

Solution

With $\gamma = 10\mathrm{dB}$, and $\gamma/\mathit{\Gamma} = 10\mathrm{dB} - 20\mathrm{dB} = -10\mathrm{dB} = 0.1$, we solve for the probability that the

SNR will drop below $10\mathrm{dB}$, as follows:

$$P(\gamma_1,\gamma_2,\gamma_3,\gamma_4 \leq 10\mathrm{dB}) = [1 - \exp(-0.1)]^4 = 8.2 \times 10^{-5}$$

or, using selection diversity, we can say that

$$P(\gamma_i > 10\mathrm{dB}) = 1 - 8.2 \times 10^{-5} = 0.9999$$

Without diversity,

$$P(\gamma_1 \leq 10\mathrm{dB}) = [1 - \exp(-0.1)]^1 = 0.095$$

$$P(\gamma_1 > 10\mathrm{dB}) = 1 - 0.095 = 0.905$$

Diversity-Combining Techniques

The most common techniques for combining diversity signals are **selection**, **feedback**, **maximal ratio**, and **equal gain**.

Selection combining used in spatial diversity systems involves the sampling of M antenna signals, and sending the largest one to the demodulator. Selection-diversity combining is relatively easy to implement but not optimal because it does not make use of all the received signals simultaneously.

With **feedback** or scanning diversity, the M signals are scanned in a fixed sequence until one is found that exceeds a given threshold. This one becomes the chosen signal until it falls below the established threshold, and the scanning process starts again. The error performance of this technique is somewhat inferior to the other methods, but feedback is quite simple to implement.

In **maximal-ratio** combining, the signals from all of the M branches are weighted according to their individual SNRs and then summed. The individual signals must be cophased before being summed.

Maximal-ratio combining produces an average SNR γ_M equal to the sum of the individual average SNRs, as shown below:

$$\gamma_M \qquad {}^{M}_{i} \; \gamma_i \qquad {}^{M}_{i} \; \mathit{\Gamma} \qquad M\mathit{\Gamma}$$

where we assume that each branch has the same average SNR given by $\gamma_i \quad \mathit{\Gamma}$.

Thus, maximal-ratio combining can produce an acceptable average SNR, even when none of the individual i γ is acceptable. It uses each of the M branches in a cophased and weighted manner such that the largest possible SNR is available at the receiver.

Equal-gain combining is similar to maximal-ratio combining except that the weights are all set to unity. The possibility of achieving an acceptable output SNR from a number of unacceptable inputs is still retained. The performance is marginally inferior to maximal ratio combining.

Modulation Types for Fading Channels

An amplitude-based signaling scheme such as **amplitude shift keying (ASK)** or **quadrature amplitude modulation (QAM)** is inherently vulnerable to performance degradation in a fading environment. Thus, for fading channels, the preferred choice for a signaling scheme is a frequency or phase-based modulation type.

In considering **orthogonal FSK** modulation for fading channels, the use of **MFSK** with M = 8 or larger is useful because its error performance is better than binary signaling. In slow **Rayleigh fading** channels, **binary DPSK** and **8-FSK** perform within 0.1 dB of each other.

In considering **PSK** modulation for fading channels, higher-order modulation alphabets perform poorly. **MPSK** with M = 8 or larger should be avoided.

Example: Phase Variations in a Mobile Communication System

The **Doppler** spread $f_d = V/\lambda$ shows that the fading rate is a direct function of velocity. **Table 1** shows the **Doppler** spread versus vehicle speed at carrier frequencies of 900 MHz and 1800 MHz. Calculate the phase variation per symbol for the case of signaling with **QPSK** modulation at the rate of 24.3 kilosymbols/s.

Assume that the carrier frequency is 1800 MHz and that the velocity of the vehicle is 50 miles/hr (80 km/hr). Repeat for a vehicle speed of 100 miles/hr.

Table 1

Velocity		Doppler (Hz)	Doppler (Hz)

miles/hr	km/hr	900 Mhz (λ = 33cm)	1800 Mhz (λ = 16.6cm)
3	5	4	8
20	32	27	54
50	60	66	132
80	108	106	212
120	192	160	320

Solution

At a velocity of 100 miles/hr:

$$\mathit{\Delta}\theta/\mathrm{symbol} = \frac{f_d}{R_t} \times 360^o$$

$$= \frac{132\mathrm{Hz}}{24.3 \times 10^3 \mathrm{symbols/s}} \times 360^o$$

$$= 2^o/\mathrm{symbol}$$

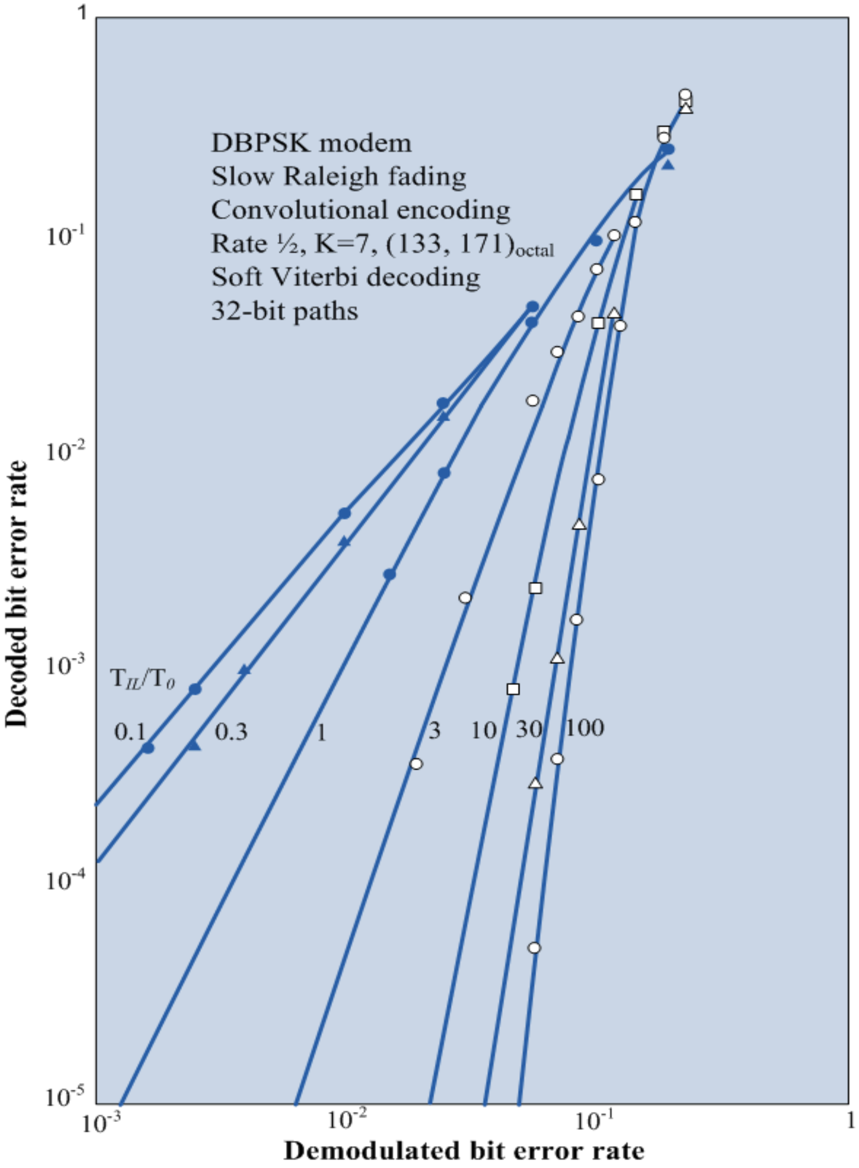
At a velocity of 100 miles/hr: $\mathit{\Delta}\theta/\mathrm{symbol} = 4^o/\mathrm{symbol}$

Thus, it should be clear why MPSK with a value of M > 4 is not generally used to transmit information in a multipath environment.

The Role of an Interleaver

The primary benefit of an interleaver for transmission in fading environment is to provide time diversity (when used along with error-correction coding).

Figure 1 illustrates the benefits of providing an interleaver time span T_{IL}, that is large compared to the channel coherence time T_0, for the case of **DBPSK** modulation with soft-decision decoding of a rate 1/2, K = 7 convolutional code, over a slow **Rayleigh-fading** channel.

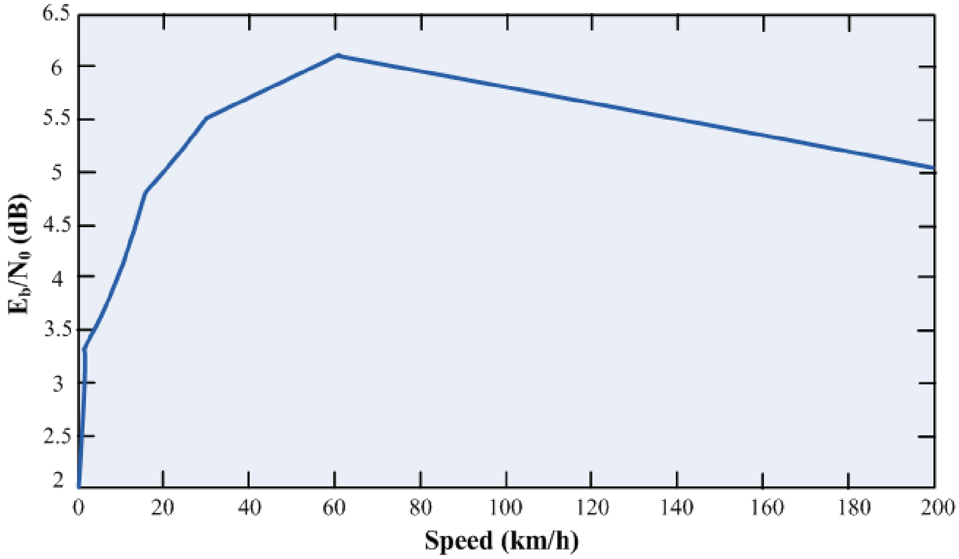


It should be apparent that an interleaver having the largest ratio of T_{IL}/T_0 is the best-performing (large demodulated BER leading to small decoded BER). This leads to the conclusion that T_{IL}/T_0 should be some large number—say 1,000 or 10,000. However, in a real-time communication system this is not possible because the inherent time delay associated with an interleaver would be excessive.

The previous section shows that for a cellular telephone system with a carrier frequency of 900 MHz, a T_{IL}/T_0 ratio of 10 is about as large as one can implement without suffering excessive delay.

Note that the interleaver provides no benefit against multipath unless there is motion between the transmitter and receiver (or motion of objects within the signal-propagating paths). The system error-performance over a fading channel typically degrades with increased speed because of the increase in Doppler spread or fading rapidity. However, the action of an interleaver in the system provides mitigation, which becomes more effective at higher speeds

Figure 2 show that communications degrade with increased speed of the mobile unit (the fading rate increases), the benefit of an interleaver is enhanced with increased speed. This is the results of field testing performed on a **CDMA** system meeting the **Interim Specification 95 (IS-95)** over a link comprising a moving vehicle and a base station.

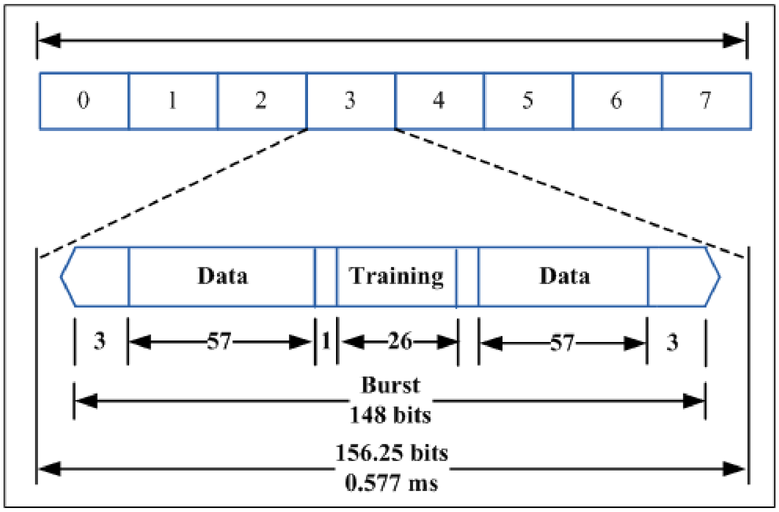


Typical E_b/N_0 performance versus vehicle speed for 850 MHz links to achieve a frame-error rate of 1 percent over a Rayleigh channel with two independent paths

.

The Viterbi Equalizer as Applied to GSM

The **GSM time-division multiple access (TDMA)** frame in **Figure 1** has duration of 4.615 ms and comprising 8 slots, one assigned to each active mobile user. A normal transmission burst occupying one time slot contains 57 message bits on each side of a 26-bit **midamble**, called a **training** or **sounding sequence**. The slot-time duration is 0.577 ms (or the slot rate is 1733 slots/s). The purpose of the midamble is to assist the receiver in estimating the impulse response of the channel adaptively (during the time duration of each 0.577 ms slot). For the technique to be effective, the fading characteristics of the channel must not change appreciably during the time interval of one slot.



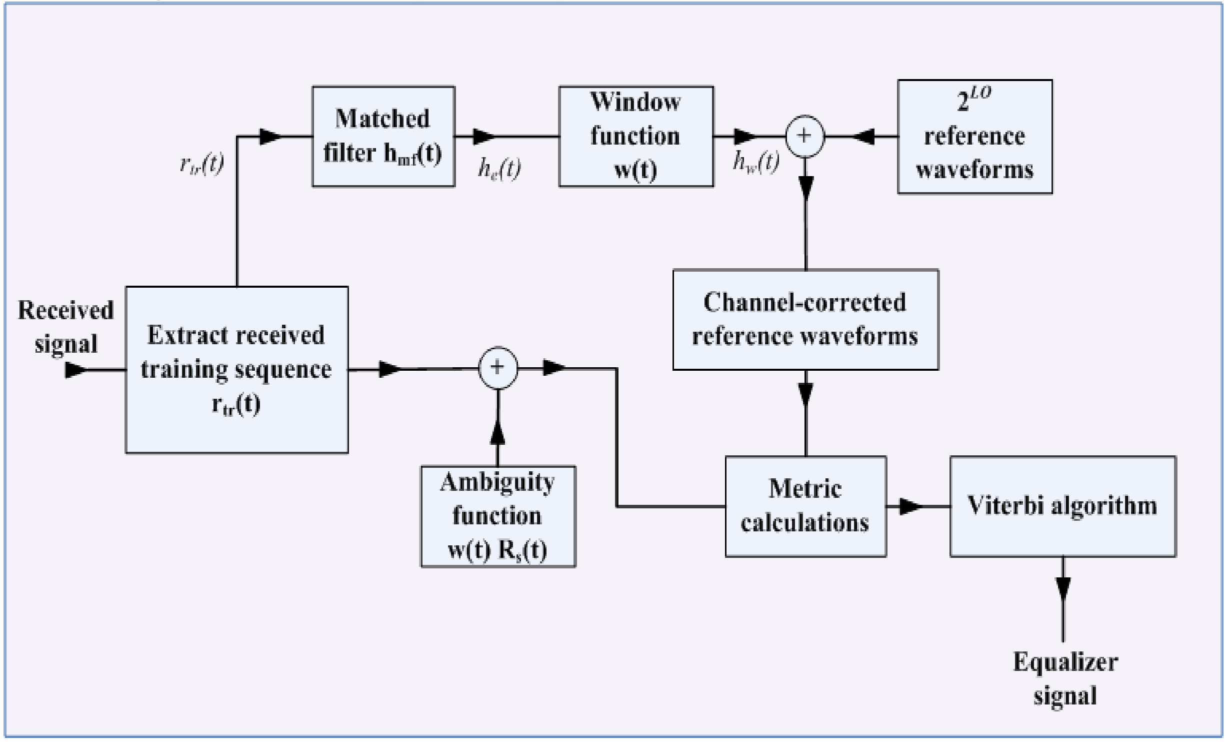
Consider a GSM receiver used aboard a high-speed train, traveling at a constant velocity of 200 km/hr (55.56 m/s). Assume the carrier frequency to be 900 MHz (the wavelength is λ = 0.33 m). The distance corresponding to a half-wavelength is traversed in $T_0 \approx \frac{\lambda/2}{V} \approx \mathbf{3}$ corresponds approximately to the coherence time. Therefore, the channel coherence time is more than five times greater than the slot time of 0.577 ms. The time needed for a significant change in channel fading characteristics is relatively long compared to the time duration of one slot.

The GSM symbol rate (or bit rate, since the modulation is binary) is 271 kilosymbols/s; the bandwidth, W, is 200 kHz. Since the typical rms delay spread σ_τ in an urban environment is on the order of 2μs, then the resulting coherence bandwidth:

$$f_0 \approx \frac{1}{5\sigma_\tau} \approx 100\text{kHz}$$

Since $f_0 < W$, the GSM receiver must utilize some form of mitigation to combat frequency-selective distortion. To accomplish this goal, the **Viterbi equalizer** is typically implemented.

Figure 2 shows the basic functional blocks used in a GSM receiver for estimating the channel impulse response.


Matched filter $h_{mf}(t)$
Window function w(t)
2^{LO} reference waveforms
$r_{tr}(t)$
$h_e(t)$
$h_w(t)$
+
Received signal
Extract received training sequence $r_{tr}(t)$
Channel-corrected reference waveforms
+
Ambiguity function w(t) $R_s(t)$
Metric calculations
Viterbi algorithm
Equalizer signal

This estimate is used to provide the detector with channel-corrected reference waveforms as explained below: (the **Viterbi algorithm** is used in the final step to compute the **MLSE** of the message bits)

Let $s_{\text{tr}}(t)$ be the transmitted midamble training sequence, and $r_{\text{tr}}(t)$ be the corresponding received midamble training sequence. We have:

$$r_{\text{tr}}(t) = s_{\text{tr}}(t) * h_c(t)$$

At the receiver, since $r_{\text{tr}}(t)$ is part of the received normal burst, it is extracted and sent to a filter having impulse response $h_{\text{mf}}(t)$, that is matched to $s_{\mathrm{tr}}(t)$. This matched filter yields at its output an estimate of $h_c(t)$, denoted $h_e(t)$:

$$h_e(t) = r_{\mathrm{tr}}(t) * h_{\mathrm{mf}}(t)$$

$$= s_{\mathrm{tr}}(t) * h_c(t) * h_{\mathrm{mf}}(t)$$

where $R_s(t) = s_{\mathrm{tr}}(t) * h_{\mathrm{mf}}(t)$ is the **autocorrelation function** of $s_{\mathrm{tr}}(t)$. If $s_{\mathrm{tr}}(t)$ is designed to have a highly-peaked (impulse-like) autocorrelation function $R_s(t)$, then $h_e(t) \approx h_c(t)$.

Next, we use a windowing function, $w(t)$, to truncate $h_e(t)$ to form a computationally affordable function, $h_w(t)$. The time duration of $w(t)$, denoted L_0, must be large enough to compensate for the effect of typical channel-induced ISI. The term L_0 consists of the sum of two contributions, namely L_{CISI}, corresponding to the controlled ISI caused by Gaussian filtering of the baseband waveform (which then modulates the carrier using MSK), and L_C, corresponding to the channel-induced ISI caused by multipath propagation. Thus,

$$L_0 = L_{\mathrm{CISI}} + L_C$$

The GSM system is required to provide distortion mitigation caused by signal dispersion having delay spreads of approximately 15–20 μs. Since in GSM the bit duration is 3.69 μs, we can express L_0 in units of bit intervals. Thus, the **Viterbi equalizer** used in GSM has a memory of 4–6 bit intervals. For each L_0-bit interval in the message, the function of the Viterbi equalizer is to find the most likely L_0-bit sequence out of the 2^{L_0} possible sequences that might have been transmitted.

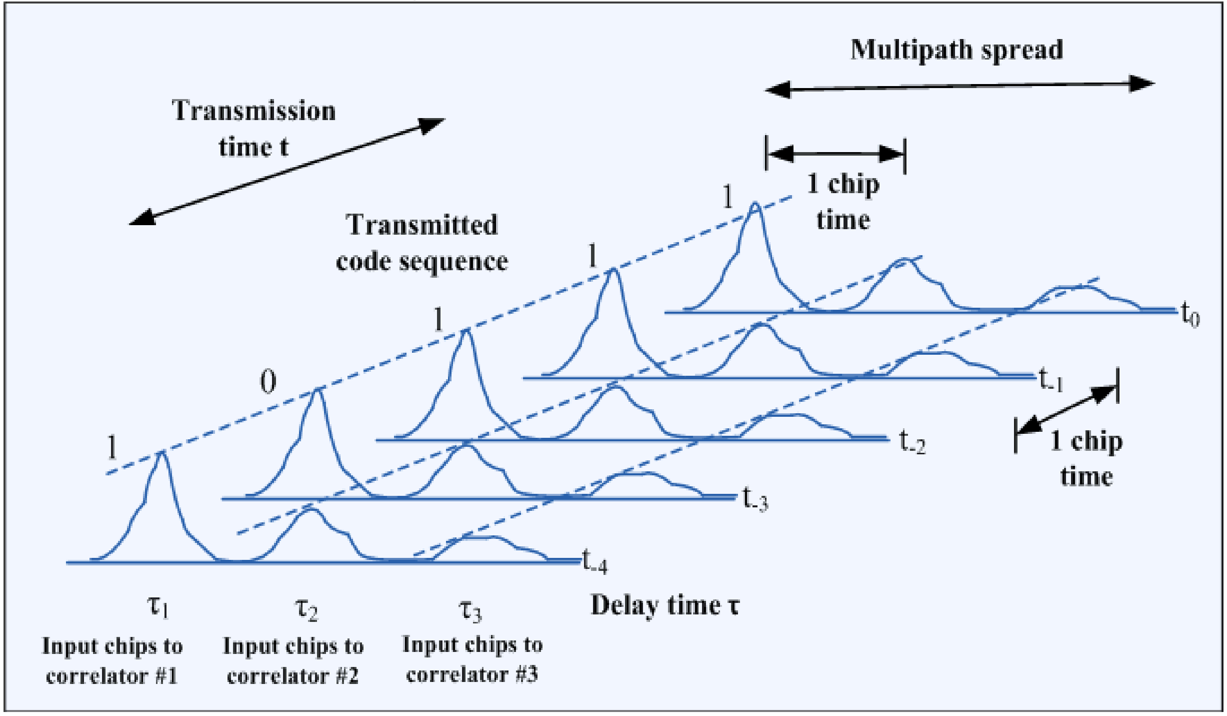
Determining the most likely transmitted L_0-bit sequence requires that 2^{L_0} meaningful reference waveforms be created by disturbing) the 2^{L_0} ideal waveforms (generated at the receiver) in the same way that the channel has disturbed the transmitted slot. Therefore, the 2^{L_0} reference waveforms are convolved with the windowed estimate of the channel impulse response, $h_w(t)$ in order to generate the disturbed or so-called channel-corrected reference waveforms.

Next, the channel-corrected reference waveforms are compared against the received data waveforms to yield metric calculations. However, before the comparison takes place, the received data waveforms are convolved with the known windowed autocorrelation function $w(t)R_s(t)$, transforming them in a manner comparable to the transformation applied to the reference waveforms. This filtered message signal is compared to all possible 2^{L_0} channel-corrected reference signals, and metrics are computed in a manner similar to that used in the **Viterbi decoding algorithm**. It yields the **maximum likelihood estimate** of the transmitted data sequence.

The Rake Receiver Applied to Direct-Sequence Spread-Spectrum (DS/SS) Systems

Interim Specification 95 (IS-95) describes a Direct-Sequence Spread-Spectrum (DS/SS) cellular system that uses a **Rake receiver** to provide path diversity for mitigating the effects of frequency-selective fading. The Rake receiver searches through the different multipath delays for code correlation and thus recovers delayed signals that are then optimally combined with the output of other independent correlators.

Figure 1 show the power profiles associated with the five chip transmissions of the code sequence 1 0 1 1 1. Each abscissa shows three components arriving with delays τ_1, τ_2, and τ_3. Assume that the intervals between the transmission times t_i and the intervals between the delay times τ_i are each one chip in duration. The component arriving at the receiver at time t_{-4}, with delay τ_3, is time-coincident with two others, namely the components arriving at times t_{-3} and t_{-2} with delays τ_2 and τ_1 respectively. Since in this example the delayed components are separated by at least one chip time, they can be resolved.



At the receiver, there must be a sounding device dedicated to estimating the τ_i delay times. Note that the fading rate in mobile radio system is relatively slow (in the order of milliseconds) or the channel coherence time large compared to the chip time duration ($T_0 > T_{\text{ch}}$). Hence, the changes in τ_i occur slowly enough that the receiver can readily adapt to them.

Once the τ_i delays are estimated, a separate correlator is dedicated to recovering each resolvable multipath component. In this example, there would be three such dedicated correlators, each one processing a delayed version of the same chip sequence 1 0 1 1 1. Each correlator receives chips with power profiles represented by the sequence of components shown along a diagonal line. For simplicity, the chips are all shown as positive signaling elements. In reality, these chips form a **pseudonoise (PN)** sequence, which of course contains both positive and negative pulses. Each correlator attempts to correlate these arriving chips with the same appropriately synchronized PN code. At the end of a symbol interval (typically there may be hundreds or thousands of chips per symbol), the outputs of the correlators are coherently combined, and a symbol detection is made.

The interference-suppression capability of DS/SS systems stems from the fact that a code sequence arriving at the receiver time-shifted by merely one chip will have very low correlation to the particular PN code with which the sequence is correlated. Therefore, any code chips that are delayed by one or more chip times will be suppressed by the correlator. The delayed chips only contribute to raising the interference level (correlation sidelobes).

The mitigation provided by the Rake receiver can be termed path diversity, since it allows the energy of a chip that arrives via multiple paths to be combined coherently. Without the Rake receiver, this energy would be transparent and therefore lost to the DS/SS receiver.